nature

SEE HOW THEY RUN

Atom-and-nucleus transport in real time

AFGHANISTAN
Natural resources
as a lifeline

CLIMATE POLITICS
A radical alternative

AUTISM BOOKS
Why memos matter...
and how science is done

NATUREJOBS
Postdocs to spare

The war on want

Many 'developing' countries are much more developed than some people think. Their rapid progress should inspire scientists and their institutions to do more to confront global poverty.

To those living in comfort in the affluent parts of the world, the stark reality behind the word poverty is difficult to comprehend. Yet it is more incomprehensible that in the twenty-first century, hundreds of millions of people are so poor that their lives amount to little more than a daily struggle to survive. Faced by such desperate circumstances, the affluent often succumb to an overwhelming sense of impotence. But defeatism would be a mistake: the battle against poverty can be won, and science can help.

Looking at the powerhouse of medical research that is the US National Institutes of Health (NIH), it is difficult to imagine that it began life 120 years ago in a small attic room in the Marine Hospital Service on Staten Island, New York. Joseph Kinyoun's one-man 'laboratory of hygiene', the first federal disease lab, was created to identify the causes of several infectious diseases that at the time were responsible for high mortality in the United States. The rest is history.

That concept of time is vital to understanding poverty. In 1915, the United States resembled India today on several poverty indicators, from child mortality to life expectancy, and Mexico has now achieved standards of human development equivalent on some scores to Sweden in the 1950s. Many of the countries labelled as 'Third World' are in fact well along the same trajectory to prosperity as today's wealthy nations. Yet growing prosperity means that new technologies are needed even more urgently, because as more countries escape the shackles of poverty, demand for energy, water and other resources will exacerbate the problems, such as climate change, that face our planet.

Exploded myths

It is essential that the myths about poverty are debunked for policymakers. Few scientists have done more in this regard than Hans Rosling, a global-health researcher at the Karolinska Institute in Stockholm. Extracting and collating vast amounts of data on poverty, he has created stunning visualizations that vividly dispel many preconceived ideas about poverty (see http://tinyurl.com/34bsyr). He also highlights the importance of factors such as good governance and institutional capacity building, showing, for instance, that although economic growth is the motor of escape from poverty, that escape is achieved faster when public health is improved.

This is an example of how science can help to frame the problems properly. Aid agencies must similarly pay more attention to evaluating the effectiveness of resource allocations and interventions. Too often in the past, advocacy of policies has been based on the latest economic theory in fashion, rather than on hard evidence. Better data on poverty indicators, in particular at the subnational level, are also needed if we are to assess the true effects of the interventions designed to meet the UN Millennium Development Goals of halving poverty by 2015 (see *Nature* **446,** 347; 2007).

In one welcome initiative, the World Bank has embraced an innovative approach to evaluation. Pioneered at the Massachusetts Institute of Technology in Cambridge, it borrows the techniques of randomized trials used in medicine to assess the impact of health and education initiatives, such as bed nets for malaria, and the factors that affect their success (see page 957).

But fighting poverty more effectively demands a change of culture not only in aid agencies but also in academia, in particular to bridge the gap between basic research and the development of drugs and vaccines for neglected diseases. The scientific community's reward structures, excessively focused on papers and patents, need to encourage efforts to use science to generate tangible benefits — not least in alleviating poverty and its impacts.

> "Leaders of academic institutions should be asking themselves how their organizations can do more to combat poverty."

Measured response

Only a few universities have, for example, created centres for biomedical translational research, evaluated not by the usual metrics but by their results, much as a company assesses project performance. Far more could be done along these lines in agriculture, energy and health.

That said, there are encouraging signs that decades of focus on wealth creation are slowly giving way to a greater social conscience. Technology-transfer offices at the NIH and many universities, for example, are increasingly waking up to the importance of making their patents freely available to groups developing technologies for use in poor countries (see *Nature* **449,** 158–159; 2007).

At a time when many sectors of society are actively engaged in working towards the Millennium Development Goals, the leaders of academic institutions should be asking themselves how their organizations can do more to combat poverty and improve global health. The eight goals themselves are ambitious to say the least. It will be a tall order to halve extreme poverty, roll back child mortality and provide universal primary education, all by 2015.

But many previously impoverished countries are now well on their way to enjoying levels of child mortality, longevity and health comparable to those of the richest countries. That should encourage scientists to help speed up the progress of those nations that still have far to go — and to strengthen their scientific infrastructure. The gleaming, state-of-the art Manhiça Health Research Center near Maputo in Mozambique, for example, may seem incongruous with the immediate disease and misery on its doorstep. Yet in a few decades it may be remembered as that nation's Staten Island lab. ∎

This article is part of the Global Theme on Poverty and Human Development, organized by the Council of Science Editors. All articles from the Nature Publishing Group are available free at www.nature.com/povhumdev. The content from all participating journals can be found at www.councilscienceeditors.org/globalthemeissue.cfm.

Watson's folly

Debate about sensitive scientific issues needs to be forthright but not crass.

So 'Honest Jim' Watson has finally fallen victim to his notorious propensity for making outrageous statements — forced to cancel a UK lecture tour and suspended from his leadership role at Cold Spring Harbor Laboratory in New York after being quoted in a British newspaper as claiming that black people are less intelligent and employable than whites (see page 960).

Watson certainly has a track record in making distasteful remarks. He has on many previous occasions voiced unpalatable views tinged with racism and sexism, ranging from a desire to see a world full of genetically engineered pretty girls to his belief that sex drive is related to skin colour. He has been largely indulged over the years, mostly in light of his towering achievement of 54 years ago in helping to deduce the structure of DNA, his ensuing Nobel prize, and his role in founding the Human Genome Project.

His latest outburst marks the point at which his views have finally been deemed beyond the pale. And rightly so — for one of the world's most high-profile scientists to state such views demonstrates a sheer unacceptable offensiveness. Watson has apologized and retracted the outburst, claiming to have been "mortified" at the outcome of the interview although he did not deny its contents. He acknowledged that there is no evidence for what he claimed about racial differences in intelligence. But the damage has been done, lending succour and comfort to racists around the globe.

Given the media storm breaking as the week progressed, it would have required remarkable bravery on the part of those hosting Watson's speaking engagements — part of the launch of his new book — to persist with them. Several high-profile establishments cancelled. This is regrettable: science is about nothing if not openness and critical debate. Scientists with controversial arguments need to be able to withstand the heat, defending or retracting statements as the evidence indicates is required. Watson, however unpleasant his utterances, has always been willing to act in this spirit.

> "The damage has been done, lending succour and comfort to racists around the globe."

The cancellations run the risk of playing into the hands of those who wish to suppress scientific inquiry. Many human geneticists are engaged in the sensitive task of unravelling differences between the world's population groups, all the while acknowledging that 'race' is an emotive and unscientific word. Others are investigating the equally sensitive genetics of 'desirable' traits, such as cognitive ability.

Asking such questions has always been controversial, given the potential for abuse of the outcomes demonstrated by the history of eugenics. Scientists explore the world as it is, rather than as they would like it to be. There will be important debates in the future as we gain a fuller understanding of the influence of genetics on human attributes and behaviour. Crass comments by Nobel laureates undermine our very ability to debate such issues, and thus damage science itself. ∎

Going it alone

The likely derailment of a US–Indian nuclear deal highlights the limitations of bilateral arrangements.

In March 2006, US President George Bush and Indian Prime Minister Manmohan Singh reached an unprecedented agreement for civilian nuclear cooperation between the two nations. In exchange for India opening up a portion of its reactors to inspection by the International Atomic Energy Agency (IAEA), the United States would agree to lift sanctions that prevent nuclear trade.

Non-proliferation advocates were critical of the move. A moratorium on nuclear cooperation with India has been in place for decades as a consequence of its failure to sign the nuclear non-proliferation treaty (NPT), which allows nations access to civilian nuclear technology as a quid pro quo for not developing nuclear weapons. India never signed the NPT and has developed nuclear weapons.

Critics argued that by arranging to sell India nuclear power technology, the United States was undermining the NPT's already shaky foundations. But the agreement was consistent with the Bush administration's penchant for pursuing bilateral, one-off deals, which its officials argue are easier to negotiate and more in line with US interests than multilateral treaties.

Supporters said that the IAEA inspections would bring India closer to the norms of the global non-proliferation regime, and that it was better for some of the country's facilities to be open to inspection than none at all.

But on 22 October, the Indian government announced that the treaty would not be approved by the end of this month, as had been hoped — raising fears that it will drift into abeyance. The Communist Party, the third largest party in the Indian parliament, has threatened to pull out of the fragile coalition government if the deal goes ahead. It opposes the deal because it would mean closer ties with the United States, possibly at the expense of relations with China. Meanwhile, the right-wing Bharatiya Janata Party has also voiced opposition, arguing that parts of the agreement would interfere with India's national defence interests.

As a result, this episode could be poised to undermine the international non-proliferation regime on two fronts. Its initiation showed that the world's largest nuclear power was ready, for the first time, to trade with nuclear states outside the NPT. But the current difficulties suggest that international inspections and other benefits to non-proliferation won't be realized from such agreements any time soon.

Bilateral arms-control deals are highly vulnerable to changes in the political climate inside either partner nation, and can undercut the legitimacy of truly international efforts. The latter have many weaknesses, but their basic permanence provides a firm foundation on which relations between nations can be built. It would be better for all involved if countries such as India and the United States eschewed bilateral fixes and got behind the further development of the existing international framework for non-proliferation. ∎

RESEARCH HIGHLIGHTS

Put the clocks back

Curr. Biol. doi:10.1016/j.cub.2007.10.025 (2007)

As a quarter of the world's population prepares for the clocks to be changed by one hour during the coming weeks, researchers claim that Daylight Saving Time really does mess you up.

Many countries move their clocks by an hour in spring and autumn to ensure more light later in the day. Studying data on 55,000 people, Till Roenneberg of Ludwig-Maximilian University in Munich and his colleagues found that the times we fall asleep and wake up — at least on free days — advance in line with the changing sunrise. But they do not adjust to the new social time imposed by daylight saving.

Their more detailed study of the sleep and activity patterns of 50 individuals in an eight-week period around each of the two annual time changes showed that all subjects adjusted rapidly to being released from daylight saving time. No-one adjusted completely to the imposition of daylight saving time, but natural early risers did better than night owls.

J. HOMA/GETTY

ENVIRONMENTAL SCIENCE
Poplar science

Proc. Natl Acad. Sci. USA doi:10.1073/pnas.0703276104 (2007)

Poplar trees are well equipped to clean up some of mankind's worst messes — if they are given a little help, that is, say researchers in the United States.

Poplar trees naturally metabolize several common pollutants, most notably trichloroethylene (TCE). By inserting into poplars a rabbit gene that encodes an enzyme — cytochrome P450 2E1 — that breaks down such chemicals in mammals, Sharon Doty at the University of Washington in Seattle and her colleagues significantly increased metabolism of TCE and other contaminants by the plants.

Transgenic poplar cuttings were able to remove 51–91% of the TCE in hydroponic solutions (pictured below) during a one-week period, compared with less than 3% uptake in the control groups. The modified poplars also remove more benzene and TCE from the air.

UNIV. WASHINGTON

CHEMICAL PHYSICS
Pearl jam

Phys. Rev. Lett. **99,** 155501 (2007)

Identical ball bearings shaken up into a random configuration will always leave at least 36% empty space between them. Why, if the packing is random, does it always stop at this value?

After decades of research, there is still no good mathematical explanation of the densest random close-packing (rcp) of spheres. But Randall Kamien and Andrea Liu of the University of Pennsylvania in Philadelphia shed new light on the matter. They propose that rcp is a metastable state — always less stable than crystalline packing, which is 10% denser. On squeezing, it runs out of ways to become still denser precisely at the observed packing fraction.

This picture might unify previous efforts that either quantified the amount of disorder in the rcp state or looked at the jamming of the particles.

CANCER BIOLOGY
Resisting hope

Cancer Cell **12,** 395–402 (2007)

A large-scale screen has identified two potential biomarkers for resistance to the breast-cancer drug trastuzumab.

Better known as Herceptin, this antibody is a powerful treatment for aggressive forms of breast cancer, but resistance is a major clinical problem.

The screen, conducted by René Bernards of the Netherlands Cancer Institute in Amsterdam and his colleagues, revealed that reduced expression of the tumour-suppressor gene *PTEN* enhances trastuzumab resistance in cultured breast-cancer cells.

PTEN inhibits a pathway mediated by the PI(3)K protein, and the researchers also found that cells that overexpress PI(3)K were trastuzumab resistant. Furthermore, analysis of 55 patients with breast cancer showed that prognosis was worse in patients who had reduced PTEN activity and an overactive PI(3)K pathway. The two genes could serve as useful markers for trastuzumab resistance, say the scientists.

CONSERVATION BIOLOGY
No home on the range

Conserv. Biol. doi:10.1111/j.1523-1739.2007.00798.x (2007)

Habitat destruction causes more species extinctions than any other factor, and species with a small geographical range are vulnerable. A new method for estimating the true ranges of forest birds has determined at what point a small range becomes too small.

Stuart Pimm and Grant Harris, then at Duke University in Durham, North Carolina, started with the published ranges of individual species. They trimmed out areas with elevations or types of vegetation that the birds avoid, as well as deforested areas. Referring to official lists of endangered birds, they then determined the range threshold that would correctly predict most birds' status. They suggest that any bird with a range of less than 11,000 square kilometres may be in trouble.
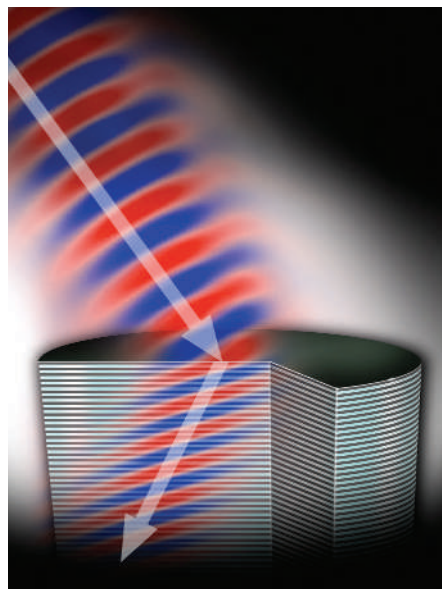
MATERIALS SCIENCE
Negative index

Nature Mater. doi:10.1038/nmat2033 (2007)

Researchers have engineered the first multilayer semiconductor able to bend incoming light in the opposite direction to all known naturally occurring materials.

Anthony Hoffman at Princeton University in New Jersey and his colleagues assembled the material by superimposing many alternating ultra-thin layers of two types of semiconductor. They found that an infrared beam shone into this at any angle was negatively refracted through the material (see picture, right).

Other such 'metamaterials' comprise alternating layers of metals and semiconductors, but can be grown to only limited thickness.

The researchers say that the new material, which is easy to produce, could improve infrared laser devices such as those used in high-speed communication and medical diagnostics, and lead to improved flat lenses to reduce distortion in telescopes and microscopes.

IMMUNOLOGY
Nod in the right direction

Immunity doi 10.1016/j.immuni.2007.08.013 (2007)

When pathogens infect a person, they prime dendritic cells to inform another, more specialized, cell of the immune system — the T cell — to tune itself up for defence. Scientists have now identified a key receptor involved in the transfer of information about a bacterial infection.

Martien Kapsenberg, Esther de Jong and their co-workers at the University of Amsterdam in the Netherlands found that

K. DRAKE

a receptor called NOD2 in the cytoplasm of dendritic cells binds to fragments of bacterial cell walls. The activated NOD2 then switches on a set of genes that results in the release of several cytokines. The particular profile of cytokines informs T cells to differentiate into Th-17 T-cells, which are tuned to recognize bacteria and organize their destruction.

Viruses, the scientists showed, did not activate the NOD2 sensor.

MICROBIAL GENOMICS
Toxic genes

Science doi:10.1126/science.1147112 (2007)

Microbes share genes between species so frequently that their evolutionary relationships aren't always clear. Edward Rubin of the Department of Energy's Joint Genome Institute in Walnut Creek,

California, and his colleagues say that there are some limits to this horizontal transfer.

Every time a microbe is sequenced, its DNA is transferred by researchers, using plasmids, into the model microbe *Escherichia coli* in what amounts to a large-scale experiment in horizontal transfer. The team took the genes that fail to transfer into *E. coli* from multiple species, and smuggled them into *E. coli* so that they would not be expressed until specifically induced. When induced, the *E. coli* stopped growing, suggesting that the genes were toxic. These transfer-resistant genes could have a future as antibiotics.

QUANTUM PHYSICS
Skipping levels

Nature Phys. doi:10.1038/nphys748 (2007)

Quantum dots are fabricated structures that confine negatively charged electrons and positively charged holes to discrete energy levels.

The rules of quantum mechanics suggest that electrons and holes added to a dot would fill the lowest available energy level before moving up to the next. But Richard Warburton of Heriot-Watt University in Edinburgh, UK, Gabriel Bester at the National Renewable Energy Laboratory in Golden, Colorado, and their colleagues have found an exception to the rule. When five or six additional positively charged holes are crammed into a dot, the researchers find that the newly added holes actually skip an energy level.

The authors say that understanding quantum dots and their unusual behaviour will be important for future applications such as quantum computing.

JOURNAL CLUB

Minhaeng Cho
Korea University, Seoul, Korea

A spectroscopist tells how the tools of his trade are revealing quantum effects in biological molecules.

In introductory quantum mechanics, one learns that particles can behave like waves, with each particle having a wavelength inversely proportional to its momentum. I am fascinated by recent work that examines how, at the molecular level, life takes advantage of these wave-like properties.

The effects aren't visible when moving biomacromolecules are viewed whole, because their large mass means they have a negligibly small 'de Broglie' wavelength. However, atoms and electrons within a molecule — for example, in active sites, where reactions such as catalysis and light-absorption take place — may interact in a wave-like way.

Researchers are thus investigating what role the wavefunctions of these molecular constituents have during biochemical reactions. And if they do interact, over what

distance and for how long does the wavefunctions' phase relationship, or quantum coherence, persist?

Spectroscopists have found evidence for coherences in a few biological systems, thanks to a technique known as multidimensional spectroscopy (A. Nagy *et al. Curr. Opin. Struct. Biol.* **16,** 654–663; 2006). This involves tracking changes in a molecule's configuration over very short timescales with laser pulses that last femtoseconds (10^{-15} s).

Further results reported this year suggest that the energy transfer in a photosynthetic system is wave-like (G. S. Engel *et al. Nature* **446,** 782–786; 2007). For this process, the quantum coherence of the light-excited charges may help the charges search out an efficient pathway through the molecule, by means of a mechanism analogous to a quantum computation.

This observation provokes a question that I look forward to seeing answered. Might biological systems have evolved to use matter's wave-like properties to optimize their efficiency?

Discuss these papers at **http://blogs.nature.com/nature/journalclub**

NEWS

Laws under review for fossils on native land

A push is underway, in the heartland of US dinosaur-hunting country, to introduce new laws to regulate the fate of non-human fossils found on Native American lands.

The drive, initiated in Nebraska, is to create legislation similar that controlling the handling and repatriation of Native American human remains, the 1990 Native American Graves Protection and Repatriation Act. Nebraska was the first state to enact a law on native human remains, even before the federal act had been passed.

On 2 November, the Lincoln-based Nebraska Commission on Indian Affairs is to consider options for a state law covering non-human fossils found on Native American lands. The commission may pass a resolution for a law, or even prepare draft legislation, says Judi Gaiashkibos, the commission's executive director.

Suspicion of scientists and their way of dealing with fossils runs deep among Native American political leaders. "Some scientists still do things deceptively," says Gaiashkibos. "To me, it is exploitation."

A complex patchwork of federal and local regulations currently governs fossils found on Native American lands. The lands themselves are a diverse collection of reservation land, land held in trust and privately owned land within a reservation. Palaeontologists routinely work with local tribes, and sometimes Native American students participate in the excavations. Nebraska already has a requirement for scientists to consult with tribes.

But such requirements don't go far enough, argues Lawrence Bradley, a doctoral geography student at the University of Nebraska-Lincoln. "In the past, the paradigm was, it was okay to go onto native lands, pull out fossils and have them in a museum for 150 years," he says. "I see myself as an ambassador for science, while creating opportunity for Native Americans."

Last week, Bradley presented a poster on what he called questionable collection practices at the annual meeting of the Society of Vertebrate Paleontology in Austin, Texas. These were mostly historical instances of tribes not being consulted by researchers before fossils were removed.

Married to a Mescalero Apache, Bradley is a passionate — some say confrontational — advocate for Native American rights. For instance, the American Museum of Natural History in New York, which has a fossil collection from the western plains dating from the 1800s, recently declined his requests for physical access to fossils after it was concluded that his correspondence seemed to be threatening to repatriate the fossils. The museum instead pointed him to online databases that it said would provide the information he requested.

Bradley is now lobbying US senators and various agencies — such as the National Park Service, which has responsibility for some Indian lands — to provide more consultation with tribal leaders before newly discovered fossils are released to scientists for study.

His campaign is now bearing some fruit. A new management plan is being considered by National Park Service regional officials for the south unit of Badlands National Park, which covers 540 square kilometres of the Pine Ridge Indian Reservation in South Dakota.

Currently, the National Park Service has an

Health official speaks out about row over drug critic

The former drug-company official who heads global health at the Bill and Melinda Gates Foundation has broken his silence over accusations of acting inappropriately eight years ago, when a physician raised questions about the safety of one of the company's drugs.

The drug, Avandia (rosiglitazone), has come under scrutiny this year after two meta-analyses[1,2] concluded that it significantly increases the risks of heart attack. So far, both US and European regulators have kept Avandia on the market, although US officials are considering regulatory action.

Fresh questions arose last month after Senator Charles Grassley (Republican, Iowa) made public a 1999 e-mail exchange between Gates official Tachi Yamada — who was then chief of research and development at Avandia's maker, SmithKline Beecham — and William Claypool, the company's director of worldwide clinical development.

Avandia had just won regulatory approval. But speaking at a scientific meeting 5 days before the exchange, John Buse of the University of North Carolina noted that there were about 50% more cardiovascular deaths in subjects who took Avandia than in those on other diabetes drugs, although the difference was not statistically significant. Buse then mis-stated the cardiovascular death rate of trial subjects taking a placebo drug, saying it was lower than in those on Avandia. He later corrected this in a statement that the company showed to investors.

But Claypool wrote to Yamada that Buse had "repeatedly and intentionally misrepresented Avandia data from the speaker's

Experts have quarrelled over Avandia's safety record.

Air permit blocks Kansas coal plants

Sunflower Electric Power Corporation had the documents in order for its plans to build a pair of 700-megawatt coal-fired power plants in Kansas, and all of the technical requirements had been met. A key state official acknowledged this last week, but refused to grant the company an air-quality permit because of concerns about global warming.

The Sunflower ruling marks the first time in the United States that power-plant plans have been halted over fears that carbon-dioxide emissions endanger people and the environment.

The decision by Roderick Bremby, Kansas secretary for health and the environment, was based in part on an April Supreme Court ruling that the US Environmental Protection Agency (EPA) has the authority to regulate carbon dioxide in automobile emissions (see *Nature* 446, 589; 2007) — and, by extension many argue, from power plants. "The secretary believes it would be irresponsible to ignore the information about the effects of carbon dioxide and greenhouse gases," says Joe Blubaugh, spokesman for the Kansas Department of Health and Environment. Sunflower blasted the decision and promised to pursue "legal and legislative remedies".

"It's a political decision," says Frank Maisano, an energy-industry lobbyist with Bracewell & Giuliani based in Washington DC. But Mark Brownstein, who handles business partnerships for green campaigner Environmental Defense, says the move is a logical extension of the Supreme Court decision.

Meanwhile, California is expected to begin proceedings to sue the EPA this week in an effort to allow the state to set its own greenhouse-gas emissions standards for automobiles.

Jeff Tollefson

Lawrence Bradley's campaign for tribal involvement has helped to revamp plans for managing Badlands National Park in South Dakota.

agreement with the reservation's Oglala Sioux under which tribal members are consulted about the fate of fossils found on their lands. But it was deemed to be insufficient, so discussions have begun with a view to including fossil discoveries in the park's new 20-year plan, says Rachel Benton, a palaeontologist with the park service. As geological formations there date to about 50 million years ago, the primary impact would be largely on mammal fossils, Benton says.

Although there was many a furrowed brow to be seen in front of Bradley's poster — which referred to "grave desecration" by pioneering scientific explorers — some palaeontologists saw advantages in his proposals. One such idea

was to create tribal museums, where fossils from Native American lands could be displayed.

"It would be a great idea for teaching natural history to kids," says Lawrence Flynn, an assistant director at Harvard University's Peabody Museum of Archaeology and Ethnology in Cambridge, Massachusetts.

Daniel Chaney, a palaeobiologist at the National Museum of Natural History in Washington DC, recalled that as a youth he worked with renowned collectors for the American Museum who would consult with tribal councils, often in the tribe's native language. "They always went to the landowner," Chaney says. ∎

Rex Dalton

dais". He suggested that Buse be warned to stop, "with the punishment being that we will complain up his academic line". He added that a lawsuit could be "reserved for a later approach".

Yamada replied: "I think that there are two courses of action. One is to sue him for knowingly defaming our product even after we have set him straight as to the facts. The other is to launch a well-planned offensive on behalf of Avandia." Yamada recommended the latter.

In an interview with *Nature* last week, Yamada said that after receiving Claypool's e-mail he called Buse's boss "to be certain that he was a legitimate

academic". He said that he was reassured to learn that Buse was "a qualified scientist making the best of the data that he had". A careful reading of the e-mails released by Grassley, Yamada added, shows that "I did not initiate any discussion of a lawsuit [against Buse]". He added: "Nor did I ever discuss a lawsuit with anybody else ... I wouldn't want the media to think it's some diabolical plot hatched by me against Dr Buse, because nothing could be further from the truth."

"I don't think that Dr Yamada is the bad guy in this story," Buse says. "He did his job and he did it in a fairly

sensitive way." He adds that he didn't have a problem, then or now, with Yamada calling his chairman to check his credentials. "That wasn't the disturbing bit. The most disturbing part was the veiled threat of a lawsuit."

Buse said that the threat came from one of many other company officials who deluged him with phone calls — prompting him to write to Yamada, asking him to "call off the dogs". ∎

Meredith Wadman

1. Nissen, S. E. & Wolski, K. *N. Engl. J. Med.* **356,** 2457–2471 (2007).
2. Singh, S. *et al. J. Am. Med. Assoc.* **298,** 1189–1195 (2007).

Sunflower Electric has been denied permission to build new coal-powered plants.

SIDELINES

Europeans forgo US labs

Tim Somervaille made a surprising move last month. After four years as a postdoc researching leukaemia at Stanford University in California, he went back home to Britain. In doing so, he joined other European biologists working in US labs who are finding that prospects in their native countries are more enticing than those in their adopted home.

Somervaille knew he would return to Britain at some point, but he hadn't expected to be heading back so soon. "The current funding landscape in Britain is, frankly, way more attractive," he says.

For decades, an abundance of positions, strong funding opportunities and more independence in the lab made the United States the place to be for ambitious European biologists. Now, stagnant budgets at the National Institutes of Health (NIH) and increased competition for grants are starting to make America less appealing. And some academics say that they are noticing more researchers heading east across the Atlantic.

"I think people are going back," says Markus Stoffel, a molecular biologist at the Swiss Federal Institute of Technology in Zurich. Stoffel last year moved from Rockefeller University in New York for family reasons. Since his return, he has seen an increase in the number of job applications to his lab from the United States and Canada. Europeans also seem more interested in staying at home after they graduate, he says: "For many young people, it's not a must any more to go to the States."

The situation varies hugely from country to country and field to field, but cancer research in Britain offers one example of how the tables may be starting to turn. Since 2002, funding for the UK National Cancer Research Institute has risen by 51% to £390 million (US$799 million). "The prospects for funding in the United Kingdom, especially in cancer, are really healthy," says Nic Jones, director of the Paterson Institute for Cancer Research in Manchester, UK, which will be Somervaille's new home.

Meanwhile, the fortunes of cancer researchers in the United States have taken a turn for the worse. As the budget of the NIH doubled in the five years up to 2003, cancer research received an influx of money and people. Since then budgets have stagnated, even declining slightly in the case of the National Cancer Institute. "I think it's at the stage of a crisis," says William Hait, president of the American Association for Cancer Research in Philadelphia, Pennsylvania.

Duncan Odom of the Cancer Research UK Cambridge Research Institute agrees. The American-born 37-year-old last year decided

CANCER RESEARCH UK

National interests: better prospects are luring European biomedics home from the United States.

to move to Britain from the Whitehead Institute for Biological Research in Cambridge, Massachusetts. Odom says he made the move partly because of the funding situation. The scarcity of US resources has left mid-career researchers scrambling for grants, he says: "They're getting their knees kicked out from under them."

By contrast, his new institute provides him with a salary and core funding for three postdocs and three graduate students. Odom says that the change has meant less time spent writing grants and more time for his research.

But the United States' long-established leadership in cancer research is far from over. Funding may have fallen, and research positions may be more elusive, says John Schiller, an oncology researcher at the National Cancer Institute in Bethesda, Maryland, but there are still huge advantages of scale and resources. "There are so many people doing so many related things here," Schiller says, adding that opportunities in the private sector, especially in the US biotechnology industry, remain unrivalled.

Even some who have come back to Europe doubt whether they are part of a wider trend. Peter Mombaerts, who is returning to Germany after 12 years at Rockefeller University, says he has yet to see evidence of a broader exodus from the United States.

But in some cases, the usual incentives of money and stability are pulling in Europe's direction for a change. The Paterson Institute, for example, offered Somervaille a six-year deal that included salaries for himself, two postdocs and a lab manager. "I don't have to write any more grants for the next five or six years," he says. "For me, the move was a no-brainer." ■
Geoff Brumfiel

Italian bioethics committee in uproar

Stem-cell researchers in Italy are worried that a shake-up of the country's bioethics advisory body could hinder a balanced representation of their ethical position on embryonic stem-cell research.

Members of the National Bioethics Committee are nominated directly by the Italian prime minister, and have in the past been predominantly Catholic. The group expresses opinions on and suggests solutions to bioethical issues, and helps prepare legislation on relevant topics.

In October 2000, the committee, which at the time included a number of scientists and independent thinkers, expressed a favourable opinion toward embryonic stem-cell research. But in 2002, then prime minister Silvio Berlusconi renewed the committee, and since then there has been increasing conflict between its Catholic and secular members. Then in 2004, the parliament approved one of the most restrictive laws worldwide on stem-cell research.

Last year, the current prime minister, Romano Prodi, streamlined the committee from 52 to 40 members, and increased the percentage of women from 25% to almost 40%. The proportion of Catholic members remained at well over 50%.

Prodi also appointed as the committee's president 76-year-old Francesco Casavola, a former president of Italy's supreme court, who is currently director of the Treccani Institute. Casavola, a moderate Catholic, appointed two secular members and one Catholic member as vice-presidents. They were Cinzia Caporale, a bioethicist and former president of the UNESCO (United Nations Educational, Scientific and Cultural Organization) Intergovernmental Bioethics Committee; Elena Cattaneo, a stem-cell researcher at the University of Milan; and Luca Marini, a lecturer in European Union Law at the University of Rome La Sapienza.

The new committee was fractious from the start. Almost half its members advocated changing the voting system from a simple majority — perhaps to a system that describes, with justifications, the range of views across the committee. This is similar to the way in which the German National Ethics Council works. "Such a procedure

Francesco Casavola has made big changes to the committee.

E. OLIVERIO/AP

encourages a constructive discussion and diminishes conflicts," says Jens Reich, former deputy chair of the German group.

On 18 September, three members of the Italian committee wrote an internal memo criticizing what they called Casavola's unilateral way of making decisions. In particular, they criticized his decision to choose members of the influential public group Science and Life, which strongly opposes research on embryonic stem cells, as committee representatives on important national and international panels. For example, the committee's representative on the commission that is revising the reproductive-technology law is the president of Science and Life.

The letter was leaked to the press, and Casavola tendered his resignation. At the end of September, Prodi asked him to stay and ratified Casavola's proposal to replace the three vice-presidents with new ones: a rabbi, a Catholic who takes secular positions, and a member of Science and Life.

On 8 October, Casavola wrote in a letter to the committee that the former vice-presidents had not adequately defended him against growing criticism. Prodi's undersecretary Giampaolo D'Andrea, defending Casavola's decision to a parliamentary commission, said that there were "inner conflicts between the vice-presidents", a point that members of the committee strongly contest.

The current situation worries Cattaneo. "Italian science needs to rise to the level of other European countries," she says, "which means also having an institution able to make decisions that can strike an appropriate balance between medical benefits and justified ethical concerns of the nation."
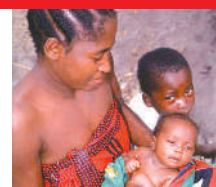
In what is likely to be a turbulent meeting on 26 October, the committee will have to produce advice about the time at which an embryo can be considered dead and could be used for research. "But with the current unbalanced representation," says Carlo Redi, a stem-cell researcher and scientific adviser at the San Matteo Hospital in Pavia, "I can hardly imagine that there could be good news for science". ■

Emiliano Feresin

955

MALARIA VACCINE HOPE FOR AFRICA'S BABIES
Mozambique trial shows vaccine safe for infants.
www.nature.com/news

Field trials aim to tackle poverty

Faced with the multitude of problems that result from and contribute to poverty, how can you decide which strategy to use to tackle an issue? One innovative lab is borrowing ideas from the medical world in a bid to find out.

The Abdul Latif Jameel Poverty Action Lab (J-PAL) is pioneering the concept of randomized trials, more commonly associated with drug safety tests, to assess what works and what doesn't in development and poverty interventions. The strategy has inspired the World Bank, which in December will choose winning proposals in a €10.4-million (US$14.9-million), 3-year programme that will use randomized trials to study the fight against poverty.

Based at the Massachusetts Institute of Technology in Cambridge, J-PAL was founded in 2003 and this year has more than 60 projects on the go in 21 countries. Esther Duflo, one of the lab's founders, says she set it up to help rigorously test the many programmes that are meant to aid the poor. "Whereas one would not dream of putting a new drug on the market without a randomized trial," she says, "such evaluations were, and to a certain extent still are, very rare for social programmes."

Although young, J-PAL has already notched up some successes. One of its first studies, involving more than 30,000 youngsters in rural Kenya, found that deworming children reduced the number of days taken off school by 25% (E. Miguel and M. Kremer *Econometrica* **72,** 159–217; 2004). Another study, in India, showed that hiring young local women to help at schools with underperforming students significantly increased test scores, and was six times cheaper than the computer-assisted learning already being tested (A. Banerjee *et al. Q. J. Econ.* **122,** 1235–1264; 2007). "J-PAL's results in education are solid and important," says Nilima Gulrajani, an expert in aid management at the London School of Economics and Political Science.

Planning trials in poor countries is very different from clinical trials, says J-PAL's executive director, Rachel Glennerster. Rather than set up its own experiments, J-PAL joins forces with groups that are already working in the field. The deworming project, for example, had already been planned. J-PAL just joined the project, which was then slightly rethought to allow the science to be done alongside.

Adding research protocols can be a hassle for development groups, says Glennerster, so it takes a lot of discussion and goodwill before

Randomized trials show local women can help improve test results in underperforming Indian schools.

J-PAL can join a project and start doing trials.

Pascaline Dupas, a researcher at Dartmouth College in Hanover, New Hampshire, and one of a network of researchers who work with J-PAL, has been tackling a more political issue: bed nets. Treated with insecticide, bed nets are one of the main new tools for controlling malaria, but debate has raged over whether widespread use is best encouraged by handing the nets out free, or charging for them to encourage responsibility. The US Agency for International Development pursues the latter policy, but on the basis of little hard evidence.

In a trial set up in Kenya, Dupas randomized the price at which pregnant women could buy nets. Her results, which are being prepared for publication, come down firmly on giving nets out free. She found that people who got free nets used them just as responsibly as those who paid for them. Moreover, charging even 75 cents reduced net use by 75%.

The lab is also exploring procrastination, which can actually be a major public-health problem. HIV testing is important for preventing the spread of AIDS, but many patients don't pick up their test results. In a soon-to-be-published study, J-PAL has found that giving people as little as 10 cents as a reward for picking up their results on the day they are ready significantly increases compliance.

The lab is currently brainstorming similar ideas to improve a major problem in tuberculosis

> **"It's the first attempt to approach poverty research in a scientific, controlled, experimental way,"**
> — Nilima Gulrajani

(TB) treatment. Patients with TB must take their drugs every day for six to eight months to eliminate infection, but often stop as soon as they feel better. With mobile phones now more common in poor countries, the researchers have come up with an idea. A text message reminds patients to take their pill. On opening the pill wrapper they get a code that gives them three minutes' free call time. "I'd love to test this in a randomized trial," says Glennerster.

Mel Spigelman, director of research and development at the Global Alliance for TB Drug Development in New York, says it has yet to work with J-PAL, but "appreciates its efforts to improve global public health". Beating TB, he says, "will take a wide range of creative efforts".

But Gulrajani urges against excessive enthusiasm for randomized trials in poverty research. She worries that policy-makers may jump on the findings as scientific too soon, and apply them too broadly — neglecting painstaking, but seemingly softer, classical social-science studies. At the same time, she praises J-PAL's concept. "It's the first attempt to approach poverty research in a scientific, controlled, experimental way," she says. "You are going to see this being increasingly adopted. It is a fantastic idea." ■

Declan Butler

See Editorial, page 947

This article is part of the Global Theme on Poverty and Human Development, organized by the Council of Science Editors. All articles from the Nature Publishing Group are available free at www.nature.com/povhumdev. The content from all participating journals can be found at www.councilscienceeditors.org/globalthemeissue.cfm.

Little consensus on egg freezing

WASHINGTON DC

Freezing a woman's eggs for use later in fertility treatments doesn't seem to greatly increase birth defects or abnormalities, new data suggest. But the findings, presented at the annual meeting of the American Society for Reproductive Medicine (ASRM) in Washington DC last week, are difficult to interpret. The field remains caught between researchers, most of whom say that the practice should still be considered experimental, and the public, which is itching to embrace the technology to extend a woman's fertile years.

Egg freezing is regarded as a positive, if costly, option for women facing illness or treatment that could truncate their fertility. But most researchers think that the procedure is not appropriate for women who wish to delay fertility for social, career or other non-health-related reasons.

"We acknowledge the interest women have in this technology," says Marc Fritz, a professor of obstetrics and gynaecology at the University of North Carolina at Chapel Hill. But "there remain very few studies in which the safety and efficacy of this practice can be judged."

Last week, an ASRM committee chaired by Fritz recommended that egg freezing continue to be considered experimental, and not be marketed or promoted to healthy young women.

Human eggs are large, fragile and filled with water, and both freezing and thawing can damage them. Freezing fertilized embryos is easier and more routine, but is less desirable for some women owing to social or ethical reasons. Some women wish to preserve eggs for the right partner, and couples can be uncomfortable with the prospect of freezing embryos.

Egg freezing, which has been in place for at least two decades, has not been used widely; doctors estimate that between 300 and 600 children have been born from formerly frozen eggs.

Safety remains one of the main concerns. At the ASRM meeting, Andrea Borini of the Centre for Reproductive Health in Bologna, Italy, reported that just 2 of 123 live births from his clinic's egg cryopreservation programme had major abnormalities: one had a nasal blockage, the other a rare developmental disorder called Rubinstein–Taybi Syndrome. (But both children had fathers with fertility problems; "we may need to look into that," says Borini.)

Meanwhile, Ilan Tur-Kaspa of the Reproductive Genetics Institute in Chicago, Illinois, pre-

Experts disagree on the safety and efficacy of egg-freezing techniques.

M. I. WALKER/WELLCOME IMAGES

sented an analysis of more than 37 publications on children born from cryopreserved eggs. Of 555 live births, including those reported by Borini's group, he turned up just 3 other major abnormalities.

Demonstrating just how difficult the data are to interpret, Kutluk Oktay from Weill Cornell Medical College in New York says that his group tallies the number of reported births from cryopreserved eggs at closer to 300, not 555. He says that the Tur-Kaspa group missed some overlap in the studies — places where researchers re-reported earlier findings or where collaborative groups reported on each others' data. But Tur-Kaspa says that they looked closely for these instances, and says that "as far as we know there's no overlap".

> "Every scrap of evidence comes with a dizzying number of caveats."

Nevertheless, the data are reassuring, Tur-Kaspa says. Five abnormalities in 550 births is a less than 1% abnormality rate — comparable to that for a natural pregnancy. The study didn't address efficacy, though. "We don't know the denominator — we don't know how many [pregnancies] were terminated," says Oktay. "It gives you a summary of how many babies are out there and how many might have an abnormality. But none of the data we've seen out there is sufficient to say the issue is resolved."

One problem in comparing studies is that eggs can be frozen several ways. The method Borini's group uses, known as slow-rate freezing, automatically cools the eggs in stages to subzero tem-

peratures, in the hope that water will be forced out of the egg before it crystallizes. Another process, vitrification, is a flash-freezing process that requires more manual involvement.

Cool tactics

Many practitioners think that vitrification is the superior technology, but few prospective studies have been done to compare the procedures. To help change that, Gary Smith, director of the Assisted Reproductive Technologies Laboratory at the University of Michigan in Ann Arbor, has been collaborating with a group at the Huntington Centre for Reproductive Medicine in São Paulo, Brazil. At the ASRM meeting, Smith presented results from a small study in which 28 slow-frozen eggs, and 35 vitrified ones, were thawed then fertilized. Further analysis of the results indicates that for every 21 oocytes frozen via vitrification, one pregnancy could be expected. The numbers were slightly worse for slow-rate freezing: 45 frozen eggs per pregnancy.

Doubts remain about egg cryopreservation, and every scrap of evidence comes with a dizzying number of caveats because of variability both in procedures and in patients. Still, some doctors are ready to broaden its use. "We're kind of at a point in this technology where it's kind of like a teeter totter," says Smith. "For me the important thing is the completion of studies like this that give the patients information about the reality of these technologies." ■

Brendan Maher

CRASHED SPACECRAFT YIELDS DATA
Scientists pick out solar-wind samples from pieces of the Genesis mission.
www.nature.com/news

USAF 388TH RANGE SQD/NASA

Security issues plague US research

A high-level 'science and security commission' is needed in the United States to limit bureaucratic restrictions on research that could thwart terrorist attacks, says a report from the National Academies.

The report's authors — an independent panel convened from the US research and security communities — agreed that universities must keep an open culture to make discoveries, even if some of those discoveries could be used by terrorists. "You need to stay ahead," says Alice Gast, president of Lehigh University in Bethlehem, Pennsylvania, and the panel's co-chair.

The group also called for an easing of visa, technology-transfer and contract restrictions — many of which still reflect the mentality of the cold war. Today, international collaboration and rapid communications render many restrictions pointless.

Research contracts are particularly troublesome, the panel found. US defence agencies often offer companies contracts that contain clauses preventing the dissemination of knowledge. That presents a problem when the companies subcontract to universities, which have an obligation to publish. The panel also found growing restrictions on research involving 'sensitive but unclassified' technology, a murky term in the view of the panel. According to a Reagan-era presidential directive, research should either be classified and restricted, or unclassified and unrestricted.

Government research-contract officers have more incentives to be cautious than open, says Tobin Smith of the Association for American Universities in Washington DC, which advocates fewer research restrictions.

"The proposed commission would push government agencies to apply more consistent rules."

"They don't want to be the ones to blame for a major security breach," he says.

The proposed commission — which would be co-chaired by the director of the Office of Science and Technology Policy (OSTP) and the president's national security adviser — would push government agencies to apply more consistent rules. Smith is pleased that the proposed commission involves the national security side as well as the the science side. Having both at the table would increase its effectiveness, he says.

Gast says she doesn't know if the commission will be created, but she is pitching the idea to the OSTP, the National Security Council and congressional committees. ■

Eric Hand

Watson suspended over comments on race

James Watson, the geneticist credited with the co-discovery of the structure of DNA, has abandoned a book tour in Britain after suggesting that he thinks black people may be less intelligent than whites.

In an interview with the London-based *Sunday Times*, published on 14 October, the 79-year-old said he was "inherently gloomy about the prospect of Africa" because "all our social policies are based on the fact that their intelligence is the same as ours — whereas all the testing says not really".

After the comments were picked up by other media on 17 October, Cold Spring Harbor Laboratory in Long Island, New York, announced that it was suspending Watson from 'administrative duties' as chancellor of the Watson School of Biological Sciences there.

Watson said he regretted his comments, but criticism from senior US colleagues was unsparing. "He has failed us in the worst possible way," said Henry Kelly, president of the Federation of American Scientists. "It is a sad and revolting way to end a remarkable career."

See Editorial, page 948.

Sutro Tower in San Francisco hosts the first of California's regional greenhouse-gas detectors.

Greenhouse-gas sensors tower over California

California has begun installing regional greenhouse-gas detection systems in metropolitan areas, becoming the first US state to gather such regional data.

Sensors have been placed on Sutro Tower in San Francisco and Richland Tower in the Sacramento suburbs as part of the California Greenhouse Gas Emissions Project, a collaboration between state and federal agencies and universities. They measure greenhouse-gas concentrations twice a day.

The project — started by researchers at the Lawrence Berkeley National Laboratory in Berkeley, California — plans eventually to monitor gases at ten locations, using the data to help establish whether California is reaching its goal of reducing atmospheric greenhouse gases by cutting state emissions. At present these total about 550 million tonnes a year.

Senate retreats from bid to ease stem-cell restrictions

US senators have pulled back from an effort to open up the rules restricting embryonic stem-cell research in the United States. On 17 October they dropped a clause that would have done so from the 2008 funding bill for the National Institutes of Health.

Senator Tom Harkin (Democrat, Iowa) says that the measure, which would have expanded the stem-cell lines available for federal funding, was dropped in a "spirit of compromise". Earlier that day, the White House had issued a policy statement declaring the provision "a transparent attempt to reverse the clear policy the President has implemented on the ethical conduct of stem cell research" and saying that Bush would veto any bill including it.

Bush's policy restricts federal funding to research on stem-cells lines developed before August, 2001; the withdrawn rider would have allowed US funding for work on scores

S. RAGAN/AP

Footprints reveal reptiles showing their age

A set of fossilized footprints found in the Grande Anse rock formation in New Brunswick, Canada, seems to nudge back the date of the earliest known reptiles. Previously, the oldest evidence of reptiles was fossil skeletons of *Hylonomus lyelli* found in Nova Scotia in 1859 and dated to about 315 million years ago. But a team led by Howard Falcon-Lang of the University of Bristol, UK, has found reptile footprints a kilometre lower in the rock strata, indicating that they are between 1 million and 3 million years older than the previous find (H. J. Falcon-Lang *et al. J. Geol. Soc.* **164,** 1113–1118 ; 2007).

<div style="text-align: right">H. FALCON-LANG</div>

of stem-cell lines derived before 15 June 2007. Harkin had attached the measure to the bill in June (see *Nature* 447, 1043; 2007).

Demand made for UK marine-science agency

A parliamentary committee has called for the establishment of a marine-science agency to oversee Britain's ocean research.

In a report released on 18 October, the House of Commons Select Committee on Science and Technology says that UK ocean research currently "lacks resources, attention, coordination and an official champion". It suggests a new agency directly responsible for public outreach, education, working with industry and coordinating ocean observation.

The committee also calls for more marine-research funding, which it says has slipped by more than 10% since 2004, to about £40 million (US$82 million) annually. Edward Hill, director of the National Oceanographic Centre in Southampton, UK, says that he agrees that the country's marine research has been fragmented and under-funded, and welcomes the report.

Jonathan Shaw, minister for marine and rural affairs, says that the government will introduce a marine bill early next year that may reflect the panel's recommendations.

Only six EU nations meet research visa deadline

The European Commission has told European Union (EU) nations to move faster to implement a directive that would make it easier for non-European scientists to work in their laboratories.

In a statement, commission vice-president Franco Frattini urged member states to pass laws to implement the directive which, among other things, allows the visiting researchers to receive special visas with a fast-track procedure (see *Nature* 437, 1215; 2005). He noted that the deadline for them to do so passed on 12 October, and warned that the commission would start infringement procedures if they failed to take prompt action.

So far only six member states — Austria, Belgium, Germany, Hungary, Portugal and Romania — have fully implemented the directive. Fifteen others are supposed to do so; the United Kingdom and Denmark, citing their special status under one of the EU treaties, have already said that they will not pass such laws.

Spending stalemate

As the battle over the US budget drags into autumn, the amount of money available for science is hostage to larger budget disputes. **David Goldston** explains.

President George W. Bush and the US Congress are at an impasse over the budget for fiscal year 2008, which began on 1 October. And no one seems to have any idea how the log-jam will be broken, or to be working particularly assiduously at finding a path forward. Congress has passed a 'continuing resolution' to keep the government running at current spending levels until 16 November; the expectation is that a second such resolution will be needed to fund the government until around Christmas — a deadline that perhaps will force some real negotiations.

In the meantime, science funding is in limbo as Congress continues to pass spending bills in line with its budget outline — which provides about \$23 billion more for discretionary spending than the president requested in February — and the president promises to veto any such bills.

Discretionary, as opposed to mandatory, spending includes money for virtually all research and development (R&D), and is the part of the budget for which Congress decides funding levels each year. Although, not counting defence, discretionary spending constitutes only about 20% of the nearly \$3-trillion federal budget, it is the subject of most of the annual political wrangling. That's because it is much harder to change mandatory programmes such as social security, which spend whatever amount is needed to meet their legal obligation to provide benefits.

It is particularly hard to guess the final outcome for science this year, because the budget dispute is more about the size of the total pot of discretionary spending than about any specific programme. Historically, the overall level of discretionary spending has been the best gauge of what level of funding science will receive. As the science-policy scholar Daniel Sarewitz pointed out in a recent article in *Issues in Science and Technology*, for the past 30 years, R&D has amounted to between 10% and 12% of total non-defence discretionary spending. If one removes spending for the Apollo space programme, the percentage has been relatively stable for at least 45 years. Congress doesn't generally build a budget around science; as a popular but relatively obscure spending area, it gets whatever Congress can afford given the parameters of the overall budget.

Even though both political parties have

PARTY OF ONE

made a point lately of portraying themselves as champions of science, science spending is still far from the most politically salient issue. The president has generally not attacked the science funding in the spending bills — not even in the case of the National Institutes of Health, for which the House of Representatives upped the president's proposed level by more than \$1 billion. But if overall discretionary spending gets cut, science may have to take part of the hit.

It's too early to tell how the budget struggle will end. So far, it has apparently been in the political interest of both sides to draw lines in the sand. The president, who has never before vetoed a spending bill, wants to revive the Republican Party's reputation for fiscal responsibility and shore up conservative support in advance of the 2008 elections. The Democrats, in control of Congress for the first time in 12 years, want to demonstrate that they will increase funding for domestic programmes — especially when they've disappointed their liberal base with their inability to reverse course in Iraq.

The critical unknown is how swing voters, who determine the election outcome, will react. Voters tend to support spending in areas the Democrats want to increase, such as education, but they may also be concerned about deficits. Moreover, neither party wants to be blamed for causing a breakdown in the budget process, and it's hard to guess who the public in the end will see as unreasonable. The facts themselves can be confusing. For example, the president argues that his budget already increases discretionary spending by \$60 billion over fiscal year 2007; the Democrats point

out that the net increases are almost exclusively in defence and homeland security.

The politics within Congress are volatile as well. Hardcore conservatives have been angry with the president for years because spending and deficits have increased on his watch, and they would no doubt vote to sustain a presidential veto. Indeed, conservatives would probably be satisfied with a year-long impasse because it would keep spending at current levels. But other Republicans may be in no mood for a veto fight or a prolonged stalemate if either strategy makes it look as if their party can't govern. That's especially true given how small the disputed amounts are compared with the overall budget. Congressman David Hobson, a middle-of-the-road Republican from the swing state of Ohio, recently told *Congress Daily*: "I don't know who's advising him up there, but the president is really out of touch. It's too little, too late for him be a fiscal conservative … he wants \$190 billion more for the war, but he's picking a fight over \$23 billion?"

With that in mind, the most effective lobbying approach for the science budget might be to join forces with other interest groups and push for higher levels of discretionary spending. But science groups never do that, fearing — probably correctly — that getting in the middle of the broader, more partisan and ideological fight over fiscal policy would just create enemies, leaving R&D spending more vulnerable.

So as this year's spending battle grinds on, scientists and their advocates are once again on the sidelines — innocent bystanders hoping to avoid becoming innocent victims. Once decisions are made on overall discretionary spending, the lobbyists will go to work trying to protect the R&D funding that Congress has provided so far. To everyone's surprise, such last-minute scrambling worked earlier this year as Congress belatedly finished its work on the fiscal year 2007 budget. This time, with more interest groups geared up to fight for money after a year of Democratic control, R&D may face more competition if the \$23-billion discretionary increase is trimmed. Lobbyists for R&D are likely to have a busy December. ∎

David Goldston is a visiting lecturer at Princeton University's Woodrow Wilson School of Public and International Affairs.

BUSINESS

Head in the clouds

'Cloud computing' is being pitched as a new nirvana for scientists drowning in data. But can it deliver? **Eric Hand** investigates.

D. YOUNG-WOLFF/ALAMY

Aim for the sky: cloud computing offers options for researchers to work with vast amounts of data.

Dennis Gannon, a computer scientist at Indiana University in Bloomington, knows all about bringing huge amounts of computer power to bear on complex scientific problems. He has at his disposal, for the purpose, Big Red, one of the world's largest supercomputers, right there on the campus.

But when Jong Youl Choi, a graduate student computer scientist at the university, approached him with a bioinformatics program that he had written, Gannon suggested they run it on Amazon's EC2 Beta program, as translating it for Big Red would be too time-consuming. Last year, the Seattle-based e-commerce firm introduced a 'cloud-computing' option that provides access to an ever-expandable 'cloud' of powerful computer servers. Gannon and Choi set up three virtual computers and uploaded their program, which seeks matches for an unknown protein sequence from a massive national database. The job took about 15 minutes and cost them less than US$2.00.

"It was easy," Gannon says. "I don't think Choi even thought about going to Big Red."

Earlier this month, Google and IBM announced their own approach to cloud computing. They say that they will offer free use of a cluster of several hundred servers to the computer-science departments of six top US research universities.

The machines will provide researchers with vast amounts of computer power for their data-crunching. Advocates of the concept say that it will fill a middle ground between the computers that most researchers currently have access to, and machines such as Big Red.

"Right now, we have no choice," says Randal Bryant, dean of computer science at Carnegie Mellon University in Pittsburgh, Pennsylvania, "except from stepping up from desktops to using these giant supercomputers."

IBM and Google say that their pilot scheme will teach computer-science students how to write software that uses the interlinked computers to scour and analyse the large data sets that back up Internet services such as e-mail, maps and social networking. In one exploratory project, for example, students developed software to scan the contents of Wikipedia for entries that might be malicious. "Students don't have experience dealing with this order of magnitude of data, or the ability to make their computer tasks work in a parallel fashion," says Dennis Quan, leader of the IBM group responsible for setting up the clusters.

The companies hope that the concept will evolve into a valuable resource for researchers who need powerful computers to crunch their data, but don't want the hassle of adapting their applications for supercomputers. They eventually plan to expand the cluster to 1,600 processors.

For now, Quan says, the firms have no plans to charge researchers for access to the cloud computers. But their investment, which has been estimated at tens of millions of dollars, may be intended to test the waters and see how badly university scientists crave more computing power. "They're probably donating this to see if there's a market there," says Richard Loft, a physicist who works with supercomputers at the National Center for Atmospheric Research in Boulder, Colorado.

On 16 October, Amazon responded to IBM and Google by expanding its program to more developers, and enhancing the service in terms of speed and storage. Their pay-as-you-go model, unveiled in August 2006, asked researchers to pay small fees of about 15 cents to store a gigabyte of data for a month, or 10 cents and upwards for an hour of processing time.

All this activity from the computer business's biggest names reflects the recent explosion in the sheer volume of data that scientists have to deal with. Rapid growth in the precision and speed of sensors and instruments, as well as in the computers themselves, has fired this expansion.

Virtual reality

Dealing with the data often falls to supercomputers such as Blue Ice, in Boulder, or Big Red. But critics say that these are more attuned to running huge models of complex systems — such as Earth's climate, or gas flow in a jet engine — than they are to the sort of intensive database mining that researchers in many disciplines now want to do. And although the centres that house them do all they can to fight the perception, supercomputers aren't seen as the most user-friendly devices around. The centres still "favour the efficiency of the machine over the efficiency of the humans that are using the machine", claims Bryant. He also says that their limitations can make using them "like the old punch-card days".

Other universities, such as the University of Texas in Austin and the University of Tokyo, have sought to address their scientists' needs by building their own supercomputing clusters. But Google, IBM and Amazon reckon that their experience in high-power applications will make it cheaper and more efficient for them to provide it as a service. If the experience of Gannon and his student is anything to go by, the cloud computer could be the workstation of the future. ■

"Students don't have experience dealing with this order of magnitude of data."
— Dennis Quan

H. PARKER-CONEY/IBM

CHILDREN OF THE REVOLUTION

M. S. Swaminathan transformed agriculture in India in the 1960s. Now **Daemon Fairless** finds him at the heart of another high-tech scheme to help the rural poor.

Monkombu Sambasivan Swaminathan has trouble making his way across a crowded conference room, not because, at 82, he walks with a slight stoop and an even slighter shuffle, but because he is intercepted at every step by a handshake and a request for a snapshot and an autograph. It is the kind of adulation normally reserved for Bollywood celebrities, not plant geneticists. But, as the father of India's Green Revolution, Swaminathan holds a revered spot in the national pantheon of public figures.

Swaminathan is credited with introducing new varieties of high-yield wheat to India during the 1960s and 1970s, catapulting the country from dependence on foreign-grain shipments to food independence within a few years. With about 60% of India's population employed in agriculture, it is difficult to overstate the significance of Swaminathan's contribution to his country. It is not unusual for former doctoral students to prostrate themselves before 'The Professor' and touch his feet — a sign of respect unparalleled in Western academia.

Despite this, Swaminathan shows no sign of resting on his laurels; instead he uses his position to further rural development in new ways. Over the past decade he has become the driving force behind a different revolution — a national movement to bring Internet and telecommunications to each of India's 600,000 rural villages. By building on existing networks of village telecom kiosks, including 80 kiosks set up by his non-profit institute, the M. S. Swaminathan Research Foundation (MSSRF),

Swaminathan hopes to create, with government and business support, enough kiosks to serve one in six villages.

Swaminathan's belief is that information and communications technologies (ICTs), if properly implemented, will help bridge India's growing urban–rural divide and forge better links between researchers and rural poor people. "Those of us working in agriculture, health and environmental sciences," he says, "have a moral and ethical responsibility to ensure that we not only do good work, but more importantly, that that work reaches the people for whom it is intended."

The growing disparity between India's urban economy, with its white-hot annual growth rate of around 9%, and its sagging rural economy yoked with massive unemployment, is of profound concern. India, with more than 1.1 billion people, remains the country with the largest number of poor people, 70% of whom live in rural areas. Moreover, the percentage of gross domestic product the government spends on rural infrastructure has been steadily declining since the late 1980s. According to the World Bank, improving the accessibility and quality of education, health care and basic infrastructure such as water, electricity, sanitation and roads are among India's biggest challenges.

How will ICT help? A significant barrier to

M. S. Swaminathan (left) wants to bring Internet technology to villages in India.

rural development in India is that, although the government has launched multiple development schemes, each designed to aid rural poor people, they are spread across so many different departments that even a seasoned bureaucrat would have trouble keeping track of them, let alone an illiterate farmer.

Basheerhamad Shadrach, Asia programme officer of a company called telecentre.org and secretary of the MSSRF's national ICT project, says that rural economic progress is hindered by a disconnect between farmers and researchers. "Agricultural extension workers," says Shadrach, "are supposed to go into the field every day and seek out the needs of farmers. But the moment they become government workers, their job is guaranteed; they simply aren't motivated enough." He says that the same is true of government employees whose job it is to provide rural health care, education and basic municipal services such as sanitation. The hope, says Shadrach, is that ICT will provide "a fresh approach" to agricultural extension, putting the information directly in the farmer's hands.

Swaminathan sees the same attitudes today that he faced during the 1960s, as a researcher with the Indian Agricultural Research Institute

RAVEENDRAN/AFP/GETTY(INSET); P. BAGLA/CORBIS

F. MOLERES/PANOS

Divided nation: children in Bangalore play computer games, but 80,000 Indian villages have no electricity.

(IARI). In India, he says, "there is a reluctance to pass on authority from the bureaucracy to the local representatives at the grass-roots level". In the mid-1960s, eager to turn India's grain crisis around, Swaminathan bypassed the orthodox procedure of growing novel varieties in a controlled environment for several years before handing the seeds over to agricultural extension workers who would, in turn, instruct the farmers on how to cultivate them. Instead, he headed straight for the farmers' fields and convinced them to become collaborators. "Some of my colleagues would say: 'why are you going to the villages? That is the duty of the extension officer.'" he recalls. "I didn't believe in that. I felt that it was my responsibility."

Between 1964 and 1966, Swaminathan, his colleagues at the IARI and local small-scale farmers planted more than 1,000 crops on farmland around Delhi and in adjoining states. The harvest was, on average, nearly 300% higher than traditional wheat varieties. "It had an electric effect on the farmers," says Swaminathan. He is hopeful that ICT will now propagate similarly throughout the country — creating a groundswell of demand for the technology.

Mixed success

This sounds good in principle, and no one doubts Swaminathan's enthusiasm or ability to secure political backing, but India's past experience with rural ICT schemes has been rife with disappointments. Ashok Jhunjhunwala, head of the Telecom and Networks Group (TeNeT) at the Indian Institute of Technology in Chennai, has been working to bring ICTs to rural India for more than 15 years. Jhunjhunwala says that the dozens of ICT projects across the country are a series of ongoing experiments, "some of which have worked", he says, but "most of which haven't".

"You'll hear about a village where ICTs have helped farmers get a better price for grain, or a village where someone has got better access to health care, but these are all anecdotal cases and don't represent the majority of ICT projects," says Jhunjhunwala. For one of its biggest projects, Jhunjhunwala's group helped to launch n-Logue, a Chennai-based company that set up about 3,500 Internet-kiosk franchises starting in 2001. "Half of them are now closed," says Jhunjhunwala, "The other half are only partially functioning." The root of the problem, he says, is that although n-Logue provides the franchisee with the equipment and training needed to run an ICT kiosk, there aren't enough services to create viable market demand. Most franchises close because they don't get adequate

return on their investment. And many kiosk operators, having received computer training, end up leaving their village to seek fortunes in the cities.

Jhunjhunwala says that n-Logue's experience is not unusual. Indeed, a 2006 study by the United Nations Development Programme of 18 ICT projects in India — representing some 6,500 ICT kiosks across 10 states — found that many faltered because they didn't address rural needs. The study found that several of the projects failed to "understand the importance of cultivating close relationships with their beneficiary community". In other words, the projects failed to listen to the villagers.

Despite these failures, Jhunjhunwala is optimistic that ICTs can eventually help. His group has had some success with a distance-education programme in 70 villages over the past 2 years. The programme more than doubled the number of rural children who passed the exams they take at 15 years old. "It was a marvellous success." But, he notes, scaling up to a national level is a different problem altogether. "This was for 60 or 70 villages," he says. "How do you make it happen in 100,000?"

This is the challenge that Swaminathan has taken on. In the late-1990s, his institute set up some of the first telecom kiosks in his home state of Tamil Nadu, with the goal of linking farmers and fishermen to the basic information they need. After a rocky start, which saw the first 4 centres close, the MSSRF's network has now grown to a total of 80 kiosks across 3 states.

In 2004, Swaminathan rallied the ICT troops, creating what he calls a National Alliance, a coalition of more than 400 organizations, including state governments and various business, academic and non-governmental organizations,

> **"We have a moral and ethical responsibility to ensure that our work reaches the people for whom it is intended."**
> — M. S. Swaminathan

with the collective aim of providing ICT access to every villager in India. Nationally, there are now more than 20,000 ICT kiosks operating in nearly all of India's 28 states, run by several dozen ICT providers, and that number is set to double by December of this year.

The alliance has been enormously successful in getting government support for ICT infrastructure. In addition to US$420 million that central and state governments are pledging towards the physical infrastructure for 100,000 telecom kiosks by 2008, roughly $850 million is also being invested to bring broadband connectivity to administrative groups of villages. They have also approved close to $565 million for the creation of state data centres as hubs for government services. It is an ambitious project considering that 80,000 villages are still without electricity and 65,000 villages have no telephone line.

People power

Swaminathan says that although he is encouraged by the government's commitment to rural ICTs, he is concerned that the plan is overly focused on providing equipment and physical infrastructure. As part of their more people-centred approach, the MSSRF established the National Virtual Academy (NVA) in 2003, as a distance-learning program for training villagers to become advocates for the ICT needs of their community. Over the past 4 years, the NVA has recruited more than 1,000 such villagers — 'NVA fellows' — each nominated by their peers and each with a track record of community service.

Kandeepan Selvarani, an NVA fellow from Embalam, a farming village with a population of 4,500, not far from the city of Puducherry, says that she is becoming recognized as a local problem-solver. Recently, she was approached by a cashew farmer who feared he would lose his harvest because his trees were losing their flowers. Selvarani visited the village telecom kiosk

Indra Gandhi (left) manages a village Internet kiosk that is popular with farmers.

— one of the first the MSSRF set up — and managed to track down an agricultural scientist who taught Selvarani how to prepare panchakavya — a traditional biopesticide comprised of cow-derived products: dung, urine and milk, as well as curd and ghee. "The farmer sprayed it on his trees," says Selvarani, "and the crop was saved."

The Embalam kiosk is a barren, concrete room with four ageing computers and a broadband connection. Yet the kiosk manager, Indra Gandhi, says that in the 8 years since it opened, her community has seen substantial changes. Before, most of the villagers were unaware of the various government support programmes for small farmers and fishermen. But now, she says, "everyone in the village knows about them". Farmers regularly come to the centre to get information on livestock management and crop diseases or pests. And the centre has helped more than 50 village cooperatives to apply for microfinancing loans.

The hub of the matter

A few kilometres away, the MSSRF Puducherry Village Resource Centre serves as a coordinating hub for eight village kiosks, including the one in Embalam, relaying the needs of villagers such as Selvarani to various experts and government institutions. Importantly, says Anburaj Thiagarajane, the centre's director, all of their activities — including a local-language community newspaper, daily weather reports and regular workshops — are determined by a collaboration between villagers, NVA Fellows and the centre's staff, who are all from local communities.

The Puducherry Centre is one of 15 such hubs run by the MSSRF, each of which links with government institutions, universities and businesses, and with one another, via a satellite connection donated by the Indian Space Research Organization in Bangalore. But focusing on the high-tech component of the project, says Thiagarajane, is missing the point. "The linkage between people is the most important part of an ICT programme, not the technology. If you really want the information to reach villages, you have to have people who are capable

Trained volunteers such as Kandeepan Selvarani become local problem-solvers.

of taking it there."

This is especially true for the most ambitious aspect of the national ICT project. Government funding for the telecom centres runs out in 4 years, so to continue operating, the kiosks must secure private investment to ensure that each kiosk will be a self-sustaining public–private outlet. Jhunjhunwala doubts that the government will achieve the national goal of building 100,000 telecom kiosks across the country by 2008, let alone achieving sustainability by 2010. So far, the government has attracted the interest of several large companies, including the Mumbai-based telecoms giant Reliance Communications, each of which has placed bids for government funds to operate several thousand kiosks.

Jhunjhunwala is sceptical that these companies will be able to make a success of it, even though there are businesses running profitable rural ICT centres. The Imperial Tobacco Company of India, one of the country's largest private corporations, selling products from cigarettes and clothing to fertilizers and pesticides, has established 6,500 kiosks that it says serve 38,500 villages in 9 states. The company built its first Internet kiosks in 2000 to buy grain directly from farmers. The kiosks provide farmers with market prices so they can decide when best to sell their harvest, and they sell directly to the company for an immediate cash payment. Imperial Tobacco says its system has reduced its procurement costs by 25–30% and claims to put more money in farmers' pockets.

The company's kiosks are not part of the National Alliance, and it is likely that in India's fast-moving telecom sector, the government-funded infrastructure will be

> **"If you really want the information to reach villages, you have to have people who are capable of taking it there."**
> — Anburaj Thiagarajane

leapfrogged by new technologies. Tata Consultancy Services in Mumbai, India's largest computer software exporter, is developing mobile-phone software for farmers. India boasts the fastest-growing mobile-phone market in the world. One-fifth of its 218 million mobile-phone users live in rural areas and the country's service providers are rapidly expanding wireless coverage to villages.

Arun Pande, who heads Tata's Innovation Labs in Mumbai, says the company has developed mobile-phone applications that give farmers a local 7-day weather forecast, pesticide and fertilizer advice and crop prices at nearby markets, in their local language. "When we talked to farmers," he says, "we realized that their questions were very simple and also extremely specific to the conditions in their field." Pande says that Tata has paired up with the MSSRF to launch trials in four farming villages in Maharashtra and Uttar Pradesh. He says the business model for such mobile-based services hasn't been developed yet but he sees them working alongside village kiosks.

Swaminathan welcomes these new initiatives. He says that he feels they complement the national movement. "Let many flowers bloom," he says, "whichever one works, great." After all, rural India still has more problems than solutions. According to O. P. Bhatt, chair of the State Bank of India, traditional banking is limited by having only 20,000 bank branches in rural areas, where 70% of India's population reside. He is keen to see the ICT kiosks used to help villagers obtain basic financial services, particularly microfinancing.

Bhatt says that unless the disparity between the rural poor and urban nouveau riche is remedied, India's economic success will be short-lived. He points to the growing influence of the Naxalite-communist groups, and recent terrorist attacks, as examples of violent responses to social inequity: "It can and will lead to social strife and political backlash," he warns. ■

Daemon Fairless is this year's winner of the IDRC-Nature fellowship.

MINE GAMES

Under the rubble of war-torn Afghanistan lie natural resources worth billions. **Rex Dalton** reports from Kabul on the scientists risking their lives to see them developed for the good of the country.

I n a canyon just outside Kabul, the rocky terrain is strewn with debris symbolizing the troubled past and tenuous future of war-torn Afghanistan.

Exploratory cores, drilled decades ago by Soviets probing for minerals, are scattered across a landscape peppered with landmines. A line of bomb craters crosses the basin, which was home to a terrorist training camp until late 2001, when US B-52s swept overhead, dropping bunker-busters in retaliation for the terrorist attacks of 11 September. Among other things, the Americans destroyed a building that had been used to store geological cores, later turned into an ammunitions dump.

Below this rubble lies a potential economic and social boon for the troubled nation — a massive copper deposit estimated to be worth US$30 billion at today's high prices. The deposit, called Aynak, has never been developed into a viable mine, but international corporations are now competing to win a major mining concession there. What

"The Taliban would have killed them if they had found the reports."
— Antony Benham

happens at Aynak could eventually serve as a model for developing Afghanistan's other natural resources, ranging from mineral wealth to reserves of coal and petroleum.

But concerns about the Aynak bidding process have set off a behind-the-scenes scramble among consulting scientists, diplomats and aid agency officials to try to ensure its success. In June, the top World Bank geological consultant to the Afghanistan government sent a report to the office of President Hamid Karzai that sharply criticized how the country's ministry of mines was handling the competition. The consultant, James Yeager, called for new analysis of the bids, more emphasis on social and economic benefits and a stronger analytical role by the inter-ministerial council that administers the process. The government is soon expected to announce two finalists for the concession, narrowing the field from the current five.

Because he raised the alarm, Yeager thinks that he was targeted for assassination. A capped

beer bottle of hydrochloric acid was slipped into the refrigerator of his heavily guarded apartment in Kabul; he stopped just short of drinking it. Yeager did not renew his World Bank contract and instead returned home to Denver, Colorado, joining numerous other consulting scientists leaving Afghanistan at a time when their experience is sorely needed. Meanwhile, researchers who remain there face a range of threats, from kidnapping to land-mines to booby traps.

Nevertheless, some are optimistic that Afghanistan's natural resources can be developed in a stable and sustainable manner. Officials at the World Bank, for instance, say that the inter-ministerial council has started to strengthen its role in the Aynak bidding process by asserting power over the ministry of mines and shifting the competition onto a steadier course. Using the phrase that has been a byword for conflict in the region since the days of Rudyard Kipling's *Kim*, World Bank mining engineer Michael Stanley says: "We are into a new phase of the Great Game."

Afghanistan is a key player in the game because of its panoply of geological riches,

created as the Indian subcontinent rams into Asia, and thrust into the air and exposed in the Hindu Kush mountain range. Coal, rare industrial metals and precious stones abound at various points along the range. The northern provinces of the country also have oil and gas reserves.

Back to work

For centuries, Aynak has been known for its copper, used for weapons, tools and trade along the Silk Road. Before withdrawing from Afghanistan in 1989, the Soviets drilled countless cores to assess the deposit, now estimated to hold 240 million tonnes of the metal. Work halted for years during Taliban rule, but after 2001, reconstruction teams started to identify the country's assets. In the United States, the Bush administration encouraged Afghan expatriates to help develop their homeland; some scientists who had fled the country returned (see 'Science after the Taliban').

Work in post-Taliban Afghanistan wasn't easy. In Kabul, the Afghanistan Geological Survey building had been reduced to a shell, pockmarked by rocket blasts. Its equipment, samples and library were destroyed; anything burnable had been used for fuel. US officials helped to reconstruct the building, spending at least $6.2 million to modernize the facilities with computers, labs, sample storage racks and a library housing old, rare and sometimes bullet-scarred reference volumes. The UK government also chipped in with US$8 million and a three-year contract for scientists from the British Geological Survey (BGS) to analyse natural resources.

Although Aynak is only about 35 kilometres southeast of Kabul, the BGS scientists were forbidden to go there without military protection. Even when they did manage to get there, they couldn't sample for minerals in the scattered cores. The cores had been blasted apart by bombs aimed at the old mining tunnels, which had been suspected of housing Osama bin Laden.

Back in Kabul, the team managed to patch together a detailed picture of the copper deposit from surviving Soviet core reports. Courageous staff from the Afghanistan Geological Survey had hidden the 20-year-old documents during the Taliban regime. "The Taliban would have killed them if they had found the reports," says Antony Benham, a mineral specialist at the BGS.

Working with these formerly hidden records, BGS scientists plugged in data from the cores to create a computerized model of copper distribution at Aynak. "It was a remarkable job," says geologist Richard Ellison, the official in

Abandoned shell casings litter the exploration tunnels around the Aynak copper deposit.

charge of the agency's contract, which ended on 1 September. The BGS is now negotiating for a new contract with the World Bank.

Power problems

The firm that eventually wins the Aynak concession will face many difficult tasks, but perhaps the most daunting will be to secure electricity for mining and smelting equipment. The villages around Aynak have only generators as a source of power, and building an enormous copper production facility will require lots of power from coal-fired plants. But before hundreds of millions of dollars are invested in power plants and mining facilities, coal supplies must be located, assessed and graded for development.

To evaluate coal resources in Afghanistan, the US Geological Survey (USGS) sent in a task force led by geologist John SanFilipo. SanFilipo had earned an international reputation for assessing coal in dangerous environments, particularly in adjacent Pakistan. There, he discovered one of Asia's largest coal reserves, the Thar deposit in the Sindh province of southeastern Pakistan. But much of Thar's coal is difficult to mine because of political difficulties in the country.

SanFilipo and his colleagues faced similar challenges in assessing Afghanistan's coal resources. The Soviets and officials from the Afghanistan Geological Survey had

previously found a massive coal band running across much of northern Afghanistan, with mining centred around the town of Pul-I-Khumri, northwest of Kabul. Another known coal band runs along the country's southeastern border with Pakistan, in the Katawaz Basin.

Historically, this coal has been tapped by artisanal methods — in mines called 'dog holes', dug by locals who use donkeys for underground hauling. With little shoring or proper ventilation, the tunnels regularly collapse, killing villagers. In these warrens, the geologists went coal-hunting, sometimes using old ways. Yeager, for instance, carried a bird to test air quality. "When the bird died, I left," he says.

Details on the locations of the coal deposit were extremely sketchy, and map conditions worse. "The Afghans had squirrelled the maps away during the Taliban days," says SanFilipo. "When I got there, they had brought them back to the Afghanistan Survey building, where they were piled like junk. We set out to organize, scan and digitize them for a permanent record."

The quality of Afghanistan's coal deposits varies greatly. Some contain coal that burns hot and clean; other coals are more problematic, rich in sulphur and fluorine, and emitting noxious gases when burnt. Just north of Aynak lies a coal deposit called Chalaw. The coal there could fire a power plant for the Aynak mine, officials say, but it is rich in fluorine — which requires added measures to limit pollution from power plants, and protective venting if burned inside houses.

Another challenge is to find coal that is not buried so deep that it can't be extracted. "The critical step is to determine where coal is not at the surface, but is still easily minable," says SanFilipo. Geologists thus search for a relatively flat area where coal is just below a weathered surface.

Dating the coal is also crucial, because coals of the same age will tend to be of the same quality. Older geological maps show coal reserves that might have been dated incorrectly. "We want to create a stratigraphic picture of coal deposits across the entire country," says SanFilipo. The USGS team uses palaeobotanical clues such as pollen to date the coal. It relies on scientists such as Rahman Ashraf, a palaeobotanist who fled Afghanistan after the Soviet invasion but now serves as special adviser to President Karzai. "It was a dream that I could return to work in my country," says Ashraf, who has also been appointed chancellor of Kabul University.

Road blocks

The power lines of Afghan politics run through nearly all attempts to characterize the country's natural resources. In one case, sources allege that SanFilipo was blocked from returning to Afghanistan for coal exploration by the actions of another USGS geologist — Afghanistan-born Said Mirzad.

Originally trained in France, Mirzad was director of the Afghanistan Geological Survey before the Soviet invasion in 1979. After that he ran computer services for a small USGS office in San Diego, California. After the terrorist attacks of 11 September, Mirzad's Afghan friendships vaulted him to the USGS headquarters in Reston, Virginia, to help coordinate resource development in Afghanistan.

Mirzad has deep and historic connections in Afghanistan, where his brother-in-law is the minister of defence. Mirzad is also the mentor of the minister of mines, Mohamad Ibrahim Adel, who was one of those criticized for the handling of the Aynak copper bidding competition. And Mirzad has powerful allies in Washington DC; both the US state and defence departments awarded him medals for outstanding service in 2005.

In Afghanistan, Mirzad has aided multiple projects, such as an airborne geological assessment he urged the Karzai government to fund after aid agencies declined. But some also see him as an obstructionist. Beginning in early 2005, SanFilipo attempted unsuccessfully to return to Afghanistan to continue his fieldwork and geological map inventory. His repeated requests to US officials in Kabul for clearance to return were denied, keeping him out of the country for 15 months. He was finally allowed to return three times in 2006, but not since then. "A geologist must go out in the field to see," says Ashraf, praising Yeager and SanFilipo's expeditions.

Sources say that denials for SanFilipo's travel to Afghanistan were traced to Mirzad, who was in Kabul advising Zalmay Khalilzad, then the US ambassador to Afghanistan. Khalilzad is arguably the Bush administration's most-favoured Afghan and has since

> **"It was a dream that I could return to work in my country."**
> — Rahman Ashraf

been appointed as the US ambassador to the United Nations. Mirzad's historic friendships also extend to the presidential palace in Afghanistan: he used to play bridge with President Karzai's father.

Mirzad, though, denies hindering SanFilipo's work in any way. "This is all gossip," he says. "There is not a shred of evidence." But neither he nor the USGS officials could explain why SanFilipo was refused access to Afghanistan during the time in question.

In October 2006, SanFilipo lectured at the annual meeting of the Geological Society of America in Philadelphia, Pennsylvania, on the poor state of mining in Afghanistan. Not long afterwards, he was removed as the project leader for the USGS effort. Since the meeting, he has declined to discuss the issue publicly.

These events set back coal exploration in Afghanistan substantially, say several sources in Afghanistan and the United States, who requested anonymity so they may continue to help the country without reprisals. "It is unforgivable what has happened, a disaster," says Mary Louise Vitelli, a US attorney in Kabul who has worked extensively in war-torn regions. "Guys like SanFilipo are rare; he produces quality analysis under difficult circumstances."

And some scientists with long-term experience in the subcontinent saw the tapping

Science after the Taliban

Daud Saba (pictured) left his native Afghanistan in search of a scientific education and peace. Now he has returned to try to help piece his strife-torn country back together.

Saba was the first Afghan to receive a doctorate in geology in India, a choice he made to avoid going to a Soviet university after the 1979 Soviet invasion. He then moved to Europe, Toronto and finally Little Rock, Arkansas, where he and his Afghan physician wife continue to advocate for their homeland.

When the Taliban were ousted after the terrorist attacks in the United States in September 2001, Saba opted to return to his country to use his environmental science skills for human development. Now he experiences his cushier life in America only during brief

vacations every few months, when he visits his wife, a transplant researcher at the University of Arkansas Medical Center.

In May, Saba was appointed as an adviser to Afghanistan's president, Hamid Karzai, specializing in human-development issues. The role came after his success in establishing a Centre for Policy and Human Development at Kabul University to promulgate proper laws. He and half a dozen colleagues work in a simple, single-story building on the sprawling campus, its tree-lined avenues alive with students. Nearby in the university's library, the United States and Iran have designated study areas side-by-side, and students move between them to study and learn.

In his new roles, Saba has led the writing of two hefty documents about development of Afghanistan. He has the respect of US and European scientists who have worked in the country for decades. Yet he says that he has been sidelined in Kabul, because of his honesty. He refuses to be quiet, juggling harrowing demands, just as do other expatriate scientists

who have returned to try to help their homeland.

When I visited, Saba's mobile phone rang constantly. One minute, there was a call from a grandson of Afghanistan's former king, Mohammad Zahir Shah — 'the father of the nation' — who was then on life support. Shortly thereafter came a desperate plea for advice from village leaders caught in fighting near Shindand, about 125 kilometres south of Saba's hometown of Herat.

"People are dying right now," Saba exclaimed, his dark eyes intent. But he was caught in a new emotional crossfire. What he could not say out loud at the time was that the fighting involved US soldiers — who were later implicated in more than 100 civilian deaths in an attempt to oust a Taliban force. **R.D.**

R. DALTON

Researchers such as Antony Benham (left) and Bob McIntosh (right) are helping to revitalize research at the Afghanistan Geological Survey building (above).

of Mirzad for a reconstruction role as counterproductive — as were other selections by the Bush administration in Afghanistan and Iraq. Jack Shroder, a geologist at the University of Nebraska in Omaha, has worked in Afghanistan for 35 years, conducting glacial, mapping and global-positioning-system studies. He has been integrally involved in the American Institute of Afghanistan Studies, a multidisciplinary organization to foster research. But Shroder says that he and his fellow institute leaders were never consulted about the Bush administration's science policy for Afghanistan. "We were the boots-on-the ground guys — in and out of Afghanistan before the terrorist attacks," he says. "They completely ignored us; they think academics are all left-wingers."

Shroder also says that he has repeatedly encountered difficulties dealing with Mirzad, whom he calls a hard-core nationalist. "He didn't want foreigners to get access to maps, even if they were helping," says Shroder. But Mirzad expressed surprise that he would be seen as an obstructionist. "I believe the only thing that can save Afghanistan is its indigenous wealth. I am completely behind that," he says.

USGS managers of international programmes, such as Asia project chief Jack Medlin, praise Mirzad for fighting to secure funds for the agency to work in Afghanistan. Even so, the USGS wanted $12 million a year for five years to develop resources in Afghanistan, but scrapes by with about $9 million a year.

On 13 November, the USGS is scheduled to release a status report on minerals in Afghanistan, after some delay. The main coal report isn't to be released until next year, albeit short of data as few USGS scientists have gone to Afghanistan this year. Expatriate Afghan geologist Shah Wali Faryad, now of Charles University in Prague, repeatedly invited USGS scientists to attend a conference on geological opportunities on 15–16 October in Kabul, but the agency didn't respond. Medlin cites security issues as the reason.

Competitive streak

As the coal debacle simmers in the background, bigger questions arise about the Aynak copper project. Nine corporations originally sought the concession, which includes an option on the nearby Darband deposit. By June, the field of contenders had been narrowed to five firms, all mining heavyweights: Strikeforce, part of Russia's largest private employer, the Basic Element Group; China Metallurgical Group, a Chinese government-owned conglomerate based in Beijing; London-based Kazakhmys Consortium, which mines and processes copper in Kazakhstan; Hunter Dickinson of Vancouver, Canada, which mines minerals internationally; and Phelps Dodge, a leading

US copper mining firm based in Phoenix, Arizona. An informed source says that a few months ago the favourite of the ministry of mines' technical group was the China Metallurgical Group, with Hunter Dickinson a distant second.

In his critique of the process, Yeager wrote that Afghan expertise wasn't being used to its fullest extent, and that officials controlled by Adel, the mining minister, had too much influence in the process. No economists, attorneys, environmentalists or foreign-affairs specialists had been involved in the technical analysis, he asserted, which violates the laws Afghanistan implemented after the Taliban were ousted. Yeager also noted the importance of the bidders' track records: the top-ranked company has come under fire for poor environmental records in mining in nations other than its native China.

Yeager also contends that the strategic implications of selecting either an Eastern or Western firm have not been addressed. If Afghanistan were to choose a Russian, Kazakh or Chinese bid, Yeager wrote, firms from Western nations might not seek other mineral concessions in the region in the future, fearing that Afghanistan's neighbours may have undue influence.

But Adel counters that the tender bids have been "very strong, and everyone is happy with the progress." He adds that he has not seen Yeager's report, but considered its transmission to Karzai's office "a breach" of the adviser's duties. "He is not directly responsible for the bidding," says Adel.

For environmental specialist Daud Saba, a human development adviser to President Karzai, the difficulties with Aynak have been particularly painful. Developing such a rich natural resource should be spearheaded by the country's leading scientists, he feels. "It breaks my heart when I see what is happening," he says. And unless Afghanistan puts resource development on a steady course, many more hearts may also be broken by the opportunities lost. ■

Rex Dalton is a US West Coast correspondent for *Nature.*

971

Old laws stop drugs being used in valuable new ways

SIR — I strongly agree with Curtis Chong and David Sullivan's Commentary calling for the development of new uses for existing medications (*Nature* **448,** 645–646; 2007). However, I don't believe that establishing a database of all the drugs known to clinical medicine would be sufficient to spur fresh applications of drugs in the developed world. The stumbling block here is not information gathering but economics and patent laws.

Frederick Goodwin, former director of the US National Institute of Mental Health, recently commented that lithium remains underused in high-risk patients — despite strong evidence that, unlike most other psychiatric medications, it significantly lowers the risk of suicide in people with manic-depressive illness (see R. J. Baldessarini *et al. Bipolar Disord.* **8,** 625–639; 2006). Similarly, although folic acid has been shown to increase the efficacy of antipsychotic medications in certain patients (see J. Levine *et al. Biol. Psychiat.* **60,** 265–269; 2006), a psychiatrist at a recent meeting told me that he could not persuade his colleagues to prescribe this comparatively harmless vitamin.

Such difficulties stem from the failure of intellectual property laws to assign sufficient value to 'use' patents, involving new uses for old medications. In effect, a discovery of immense human value — preventing suicide — is assigned a negligible economic value that fails to motivate the pharmaceutical industry to develop the drug. The antiquated laws governing new uses for existing medications need to be reformed if life-saving discoveries are to be exploited.

S. Paul Berger
Department of Veterans Affairs Medical Center, 3710 SW US Veterans Hospital Road, Mail Code R&D 64, Portland, Oregon 97239, USA

Transparent frogs show potential of natural world

SIR — Your News in Brief story 'See-through frog offers inside information' (*Nature* **449,** 521; 2007) reporting the creation of a transparent frog (*Rana japonica*) by M. Sumida and collaborators highlights an important achievement. But there are natural transparent tetrapods, which are not the result of genetic manipulation. The roughly 150 species of anurans commonly known as glass frogs (family Centrolenidae) show a high variation in degree of transparency, not only in their skin but also in the different peritonea covering their organs (see picture).

These natural 'see-through' frogs allow the observation *in vivo* of internal organs, which

Naturally transparent: ventral view of a gravid female glass frog *Hyalinobatrachium bergeri*; snout–vent length 23 mm.

has potential for useful applications in the biomedical sciences. Further, although the transparent mutants of *R. japonica* are difficult to generate (only one out of 16 is transparent) and have very low fitness, glass frogs represent a highly diverse group widespread throughout the American tropics.

We know very little about these arboreal transparent frogs. But their potential use in biomedical research highlights the importance that should be attached to the exploration of biodiversity — as stressed in your 15 March 2007 issue, dedicated to Linnaeus's legacy.

Santiago Castroviejo-Fisher*, Ignacio De la Riva†, Carles Vilà*
*Department of Evolutionary Biology, Uppsala University, Norbyvägen 18D, 75236 Uppsala, Sweden
†Museo Nacional de Ciencias Naturales/CSIC, 28006 Madrid, Spain

Mars needs technology designed for sample return

SIR — In your Q&A 'Space, science and the bottom line' (*Nature* **448,** 978; 2007), Alan Stern, NASA's associate administrator for science, expresses his determination to make the Mars sample-return programme a reality. On NASA's wish list for decades, the mission would collect and return to Earth samples of martian rocks, soils and atmosphere.

But Stern's goal won't be achieved by assembling rocket propulsion systems from currently proven satellite components, as is customary for planetary spacecraft.

There is no equivalent in the satellite world, nor in the defence industry, for a launch vehicle the size of a person, which is what's needed to return martian samples to Earth affordably. The velocity and acceleration required to lift a vehicle from Mars are twice the values needed to launch

from our Moon. Indeed, this technical challenge is so daunting that it's impossible to tell whether a Mars sample-return mission will work without actually building and testing miniature launch-vehicle technology. Such technology won't result from commercial market forces, as do other advances that benefit spacecraft, such as computer chips, lightweight imagers and even satellite propulsion.

Stern should aim to bring about the Mars sample-return mission by building and testing technology designed for the purpose, not by revisiting mission design options, which may only extend 35 years of hopes to 40 or 50.

John Whitehead
Physical Sciences Directorate, Lawrence Livermore National Laboratory, L-072 PO Box 808, Livermore, California 94551, USA

Call for scientists to speak up for human rights

SIR — As chair of the American Physical Society's Committee on International Freedom of Scientists, I wish to express my alarm about the kafkaesque situation confronting the Russian biologist Oleg Mediannikov, reported in your News story 'Russian scientists see red over clampdown' (*Nature* **449,** 122–123; 2007).

The American Physical Society is the biggest organization of physicists in the United States, with a large international membership. The society is independent of any government, and its international freedom committee is responsible for monitoring the human rights of scientists throughout the world, including the United States, and assisting those in need. In the past couple of years our efforts have been partially focused on Russia, where several scientists collaborating with foreign researchers have been intimidated and prosecuted.

As well as Mediannikov, we are actively assisting Valentin Danilov, Oscar Kaibyshev, Igor Sutyagin and Oleg Korobeinichev. We are committed to defending our colleagues at risk and we will continue to raise our voices whenever scholars encounter a prohibition on travelling, publishing or pursuing research. We call upon scientific human-rights organizations to join us to protest at restrictions, by any government, on the free exchange of ideas, be it by denying visas or barring scholars from attending conferences or taking academic positions.

Juan C. Gallardo
Brookhaven National Laboratory, Upton, New York 11973, USA

Contributions to this page may be sent to correspondence@nature.com. We welcome comments on publishing issues at Nautilus (http://blogs.nature.com/nautilus).

I. DE LA RIVA

COMMENTARY

Time to ditch Kyoto

Climate policy after 2012, when the Kyoto treaty expires, needs a radical rethink. More of the same won't do, argue **Gwyn Prins** and **Steve Rayner**.

The Kyoto Protocol is a symbolically important expression of governments' concern about climate change. But as an instrument for achieving emissions reductions, it has failed[1]. It has produced no demonstrable reductions in emissions or even in anticipated emissions growth. And it pays no more than token attention to the needs of societies to adapt to existing climate change. The impending United Nations Climate Change Conference being held in Bali in December — to decide international policy after 2012 — needs to radically rethink climate policy.

Kyoto's supporters often blame non-signatory governments, especially the United States and Australia, for its woes. But the Kyoto Protocol was always the wrong tool for the nature of the job. Kyoto was constructed by quickly borrowing from past treaty regimes dealing with stratospheric ozone depletion, acid rain from sulphur emissions and nuclear weapons. Drawing on these plausible but partial analogies, Kyoto's architects assumed that climate change would be best attacked directly through global emissions controls, treating tonnes of carbon dioxide like stockpiles of nuclear weapons to be reduced via mutually verifiable targets and timetables. Unfortunately, this borrowing simply failed to accommodate the complexity of the climate-change issue[2].

Kyoto has failed in several ways, not just in its lack of success in slowing global warming, but also because it has stifled discussion of alternative policy approaches that could both combat climate change and adapt to its unavoidable consequences. As Kyoto became a litmus test of political correctness, those who were concerned about climate change, but sceptical of the top-down approach adopted by the protocol were sternly admonished that "Kyoto is the only game in town". We are anxious that the same mistake is not repeated in the current round of negotiations.

Already, in the post-Kyoto discussions, we are witnessing that well-documented human response to failure, especially where political or emotional capital is involved, which is to insist on more of what is not working: in this case more stringent targets and timetables, involving more countries. The next round of negotiations needs to open up new approaches, not to close them down as Kyoto did.

Economic theory recognizes the futility of throwing good money after bad. In politics, however, sunk costs are often seen as political capital or as an investment of reputation and status. So we acknowledge that those advocating the Kyoto regime will be reluctant to embrace alternatives because it means admitting that their chosen climate policy has and will continue to fail. But the rational thing to do in the face of a bad investment is to cut your losses and try something different.

No silver bullet

Influenced by three major policy initiatives of the 1980s, the Kyoto strategy is elegant but misguided. Ozone depletion, acid rain and nuclear arms control are difficult problems, but compared to climate change they are relatively simple. Ozone depletion could be prevented by controlling a small suite of artificial gases, for which technical substitutes could be found. Acid rain was mainly caused by a single activity in a single industrial sector (power generation) and nuclear arms reductions were achieved by governments agreeing to a timetable for mutually verifiable reductions in warheads. None of this applies to global warming.

In practice, Kyoto depends on the top-down creation of a global market in carbon dioxide by allowing countries to buy and sell their agreed allowances of emissions. But there is little sign of a stable global carbon price emerging in the next 5–10 years. Even if such a price were to be established, it is likely to be modest — sufficient only to stimulate efficiency gains[3]. Without a significant increase in publicly funded research and

> **"The Kyoto Protocol was always the wrong tool for the nature of the job."**

B. MELLOR

development (R&D) for clean energy technology and changes to innovation policies, there will be considerable delay before innovation catches up with this modest price signal.

On present trends, for another 20 years, the world will continue installing carbon-intensive infrastructure, such as coal power plants, with a 50-year lifetime. If climate change is as serious a threat to planetary well-being as we have long believed it to be, it is time to interrupt this cycle.

Climate change is not amenable to an elegant solution because it is not a discrete problem. It is better understood as a symptom of a particular development path and its globally interlaced supply-system of fossil energy. Together they form a complex nexus of mutually reinforcing, intertwined patterns of human behaviour, physical materials and the resulting technology. It is impossible to change such complex systems in desired ways by focusing on just one thing.

Social scientists understand how path-dependent systems arise[4]; but no one has yet determined how to deliberately unlock them. When change does occur it is usually initiated by quite unexpected factors. When single-shot solutions such as Kyoto are attempted, they often produce quite unintended, often negative consequences. The many loopholes that have enabled profiteers to make money from the Clean Development Mechanism, without materially affecting emissions, are examples[5]. Therefore, there can be no silver bullet — in this case the top-down creation of a global carbon market — to bring about the desired end.

But could there be silver buckshot? Could we assemble a portfolio of approaches that would move us in the right direction, even though we cannot predict which specific ones might stimulate the necessary fundamental change? If so, what would such a portfolio look like? We believe that a radical rethink of climate policy should possess at least five central elements.

Focus mitigation efforts on the big emitters

The notion that emissions mitigation is a global commons problem, requiring consensus among more than 170 countries, lies at the heart of the Kyoto approach. Engaging all of the world's governments has the ring of idealistic symmetry (matching global threat with universal response), but the more parties there are to any negotiation, the lower the common denominator for agreement — as has been the case under Kyoto.

The G8+5 Climate Change Dialogue, established in 2006 to convene the leaders of the top 13 polluters, was a belated recognition of the error of involving too many parties, each with dramatically different stakes and agendas. In September, the United States convened the top 16 polluters. Such initiatives are summarily dismissed by Kyoto's true believers, who see them as diversions rather than necessary first

The treaty negotiated at Kyoto in 1997 has been ratified by 172 nations. Has it made a difference?

steps. However, these approaches begin to recognize the reality that fewer than 20 countries are responsible for about 80% of the world's emissions. In the early stages of emissions mitigation policy, the other 150 countries only get in the way.

Allow genuine emissions markets to evolve from the bottom up

Theoretically, the simplest way to establish a price signal would be through a carbon tax. However, past experience with Britain's fuel price escalator (1993–99) and US President Bill Clinton's attempt to introduce a modest 4.3-cent-per-gallon hike in gasoline tax, shows there are serious political obstacles to establishing a level of tax sufficiently high to encourage energy efficiency, let alone to stimulate serious investments in innovation.

An alternative price-based approach to market failure is cap-and-trade. But to work, such schemes must be built — like all genuine markets — from the bottom up. The cap shapes the market by signalling the social goal as simply as possibly: in this case, reduction of anthropogenic impact on the environment. The market does the rest. But, in trying to introduce, from the top down, a global market in all greenhouse gases and all sources and sinks, the Kyoto approach tries to do too much, too soon, especially in the absence of binding legal frameworks to enforce contracts among parties who are not bound by other strong ties.

There is no precedent for imposing a global trading system from above. First, lessons need to be learned from regional experiments with trading. The European Union Emission Trading Scheme confined itself to trading only in

carbon, but then made the fatal error of allowing governments unrestricted powers to allocate allowances instead of auctioning a limited supply, leading to a collapse in the price. The Chicago Climate Exchange is successfully trading a basket of gases, but participation is voluntary. Eventually, different trading systems could evolve and link up as sensible standard practices emerge, giving rise to a global market. But in the final analysis, cap-and-trade cannot deliver the escape velocity required to get investment in technological innovation into orbit, in time. That calls for something else.

Put public investment in energy R&D on a wartime footing

We stare at stark divergences of trends. On the one hand, the International Energy Agency predicts a doubling of global energy demand from present levels in the next 25 years. On the other, since 1980 there has been a worldwide reduction of 40% in government budgets for energy R&D[6]. Without huge investment in R&D, the technologies upon which a viable emissions reduction strategy depends will not be available in time to disrupt a new cycle of carbon-intensive infrastructure.

So investment in energy R&D should be placed on a wartime footing. This is a cause that embraces the political spectrum, including Kyoto supporters. In 1992 former US Vice-President Al Gore called for a 'strategic environment initiative' as part of his vision for a 'global Marshall Plan'. The conservative American Enterprise Institute in Washington DC also supports primary research on sustainable new energy technologies. In 2006, Lord Rees, the president of Britain's Royal Society suggested that major public investment in R&D should be kick-started by a global investment in energy technologies research on the scale of the Manhattan Project[7].

It seems reasonable to expect the world's leading economies and emitters to devote as

much money to this challenge as they currently spend on military research — in the case of the United States, about $80 billion per year. Such investment would provide a more promising foundation for decarbonization of the global energy system than the current approach.

Increase spending on adaptation

For the best part of a decade, discussion of adaptation was regarded by most participants in climate policy-making as tantamount to betrayal[8]. Even though it was widely recognized by the end of the 1980s that the existing stock of atmospheric greenhouse gases was likely to lead to some inevitable warming, the policy community suppressed discussion of adaptation out of fear that it would blunt the arguments for greenhouse-gas mitigation.

Today, although the taboo on discussing adaptation is lifting, the allocation of effort remains skewed. The (unmet) commitment of international resources to the multilateral Adaptation Fund under the United Nations Climate Change Convention is $1.5 billion, derived in part from a tax on the Clean Development Mechanism. Funds for mitigation, however, come from many sources and total at least $19 billion. We believe that global adaptation efforts need to be funded at comparable scales to those we advocate for investment in technology R&D.

Many climate activists seem to assume that slowing greenhouse-gas emissions has logical and ethical priority over adapting to climate impacts. But the ethical issues cut both ways. Current emissions reductions will mainly benefit future generations, whereas the momentum already in the climate system drives the near-term. Faced with imminent warming, adaptation has a faster response time, a closer coupling with innovation and incentive structures, and thereby confers more protection more quickly to more people. It is not clear to us that the interests of millions of people in poorer countries who depend on marginal ecosystems are best served by an exclusive preoccupation

with mitigation. Indeed, such a narrow focus is likely to be a fatal error. Mitigation and adaptation must go hand in hand.

Work the problem at appropriate scales

Climate change is a multi-level governance problem. Some commentators recognized early on that it is not just, or even primarily, a matter for negotiation among nation states[9]. However, national governments have been slow to recognize this. Global responses to climate policy can learn from the US system of federalism that encourages small-scale policy experiments at the state or local-government levels as well as with the philanthropic and private sectors. When state or local policies succeed, such experiments can be implemented at the federal level, and often with the enthusiastic support of national politicians.

David Victor at the Council on Foreign Relations and his colleagues have proposed exactly this approach to climate policy, suggesting that a "global federalism of climate policy" is emerging from the rubble of the Kyoto Protocol. Rather than the top-down universalism embodied in Kyoto, countries would choose policies that suit their particular circumstances. Ironically, this 'policies and measures' approach was being pursued before the emergence of the Kyoto regime[10]. However, it has been largely neglected in the post-Kyoto process. Although a bottom-up approach may seem painfully slow and sprawling, it may be the only way to build credible institutions that markets endorse. The agenda for the Bali conference should focus on this and on the scale-up of energy R&D rather than on drafting a 'bigger and better' version of Kyoto.

The silver buckshot approach

Sometimes the best line of attack is not head-on. Indirect measures can deliver much more: these range from informational instruments,

> "The rational thing to do in the face of a bad investment is to cut your losses and try something different."

such as labelling of consumer products; market instruments, such as emissions trading; and market stimuli, such as procurement programmes for clean technologies; to a few command-and-control mechanisms, such as technology standards[11]. The benefit of this approach is that it focuses on what governments, firms and households actually do to reduce their emissions, in contrast to the directive target setting that has characterized international discussions since the late 1980s.

Because no one can know beforehand the exact consequences of any portfolio of policy measures, with a bottom-up approach, governments would focus on navigation, on maintaining course and momentum towards the goal of fundamental technological change, rather than on compliance with precise targets for emissions reductions. The flexibility of this inelegant approach would allow early mitigation efforts to serve as policy experiments from which lessons could be learned about what works, when and where[12]. Thus cooperation, competition and control could all be brought to bear on the problem.

Does the Kyoto bandwagon have too much political momentum? We hope not. It will take courage for a policy community that has invested much in boosting Kyoto to radically rethink climate policy and adopt a bottom-up 'social learning' approach. But finding a face-saving way to do so is imperative. Not least, this is because today there is strong public support for climate action; but continued policy failure 'spun' as a story of success could lead to public withdrawal of trust and consent for action, whatever form it takes. ■

Gwyn Prins is at the London School of Economics Mackinder Centre for the Study of Long Wave Events, London WC2A 2AE, UK. Steve Rayner is at the James Martin Institute for Science and Civilization, University of Oxford, Oxford OX1 1HP, UK.

1. Victor, D. *The Collapse of the Kyoto Protocol and the Struggle to Slow Global Warming* (Princeton Univ. Press, New Jersey, 2001).
2. Prins, G. & Rayner, S. *The Wrong Trousers. Radically Rethinking Climate Policy* Joint Working Paper (James Martin Inst./Mackinder Centre, in the press).
3. British Petroleum *Statistical Review of World Energy June 2007*; www.bp.com/statisticalreview
4. Arthur, W. B. *Econ. J.* **99,** 116–131 (1989).
5. Wara, M. *Nature* **445,** 595–596 (2007).
6. Rayner, S. *The International Challenge of Climate Change: UK Leadership in the G8 and EU*, Memorandum to the Environmental Audit Committee, House of Commons (24 November 2004).
7. Rees, M. *Science* **313,** 591 (2006).
8. Pielke, R. A., Prins, G., Rayner, S. & Sarewitz, D. *Nature* **445,** 597–598 (2007).
9. Rayner, S. & Malone, E. L. *Ten Suggestions for Policy Makers* in *Human Choice and Climate Change: An International Assessment,* Vol. 4, *What Have We Learned?* (eds Rayner, S. & Malone, E.) 109–138 (Battelle Press, Columbus, Ohio, 1998).
10. Victor, D. G., House, J. C. & Joy, S. *Science* **311,** 336 (2006).
11. US Department of Energy *A Compendium of Options for Government Policy to Encourage Private Sector Responses to Potential Climate Change. Report to the Congress of the United States* (1989).
12. Verweij, M. *et al. Public Administration* **84,** 817–843 (2006).

Climate villains? Protesters have called for the United States and Australia to ratify the Kyoto Protocol.

AUTUMN BOOKS

Harmony of the hemispheres

Our brains seem to be finely tuned to music, but of what use are our musical powers and passions?

Musicophilia: Tales of Music and the Brain
by Oliver Sacks
Knopf/Picador: 2007. 400 pp. $26/£17.99

**This is Your Brain on Music:
Understanding a Human Obsession**
by Daniel Levitin
Atlantic: 2007. 320 pp. £17.99

Laura Garwin

Think of a favourite piece of music — a pop song or classical piece that you've heard hundreds of times. Think about how it starts. When you can hear the tune in your head, sing, hum or whistle it (unless you're in a library, in which case you might want to try this later). According to experiments done by Daniel Levitin and Perry Cook in the early 1990s, even if you have had no musical training, your rendition of the tune will probably be very close to the original in tempo, and — perhaps more surprisingly — also quite accurate in absolute pitch.

Why should our brains be able to perform such a feat? Of what use are our musical powers and passions? And what can they tell us about how the brain works, or how — sometimes spectacularly — it doesn't?

Oliver Sacks, continuing in the tradition of *The Man Who Mistook His Wife for a Hat* and *An Anthropologist on Mars*, addresses these questions by offering a collection of 'tales' in *Musicophilia*, illustrating yet more ways in which our brains can take us by surprise. In *This is Your Brain on Music*, Daniel Levitin presents a more systematic account of what cognitive neuroscience has to say about how we process and respond to music. Both authors make the case that music stimulates our nervous systems in unique ways, which

account for its special role in our lives, and in those of our ancestors.

Sacks is the consummate storyteller, and his extensive network of patients, friends and correspondents — supplemented by a magpie-like erudition — keeps him well supplied with raw material. A man is struck by lightning

and develops an irresistible desire to play the piano. A woman suffers seizures brought on by listening to Neapolitan songs. A man who cannot dress himself or remember what he did for a living still knows the baritone parts of hundreds of songs, performs successfully in public, and seems to recover his essential 'self' while he is singing.

In all, Sacks recounts the experiences of more than a hundred individuals, interlarding their stories with discussions of the likely neural underpinnings of the described behaviours. In a chapter on musical hallucinations, Sacks tells several hair-raising stories of frighteningly loud, incessant 'jukeboxes' playing in people's brains, and observes that these hallucinations are often 'release phenomena', brought on by a deficit of input to the auditory cortex as a result of hearing impairment, brain damage, the effects of drugs, or even a prolonged period in quiet surroundings. The idea — first suggested by Jerzy Konorski in 1967, and now supported by brain-imaging studies — is that hallucinations are generated by the activation of 'retro' connections from the brain to the sense organs. This process is normally inhibited by the presence of external sensory input.

This view of the brain as a dynamic system, finely balanced between inhibition and excitation, returns throughout the book. Musical or other artistic talents, facilitated by activity in the brain's right hemisphere, may suddenly emerge after damage to parts of the left hemisphere that normally suppress right-hemisphere functions, as happens in certain types of dementia. The potential for cross-inhibition between hemispheres — and the ability of regions of the right hemisphere to take over language
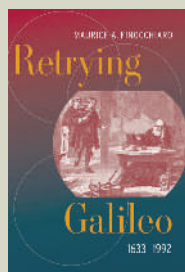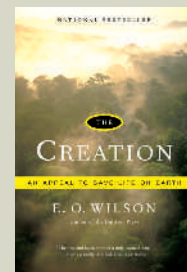

NEW IN PAPERBACK

Retrying Galileo, 1633–1992
by Maurice A. Finocchiaro (Univ. California Press, $27.50, £16.95)
The Roman Catholic Church's condemnation of Galileo in 1633 had enormous implications for both science and religion. This book explores the trial, its aftermath and its ramifications up to 1992, when the Pope expressed regret for the Galileo affair.

The Creation: An Appeal to Save Life on Earth
by E. O. Wilson (W. W. Norton, £8.99)
Ecologist E. O. Wilson rallies religious communities to join forces with scientists to save the world from environmental catastrophe, and proposes a drive to document the planet's biodiversity. Simon Conway Morris found the attempt deeply problematic, but conceded: "This clarion call, from one of the world's leading naturalists … must command respect" (*Nature* **443,** 273; 2006).

functions — also seems to explain why music therapy helps patients who have lost the ability to speak.

Many of Sacks's stories portray the striking independence of different cognitive abilities. We know of people who are tone-deaf; others may be 'rhythm-deaf', or have 'dysharmonia' — the inability to sense the harmonies created by multiple voices or instruments. After suffering brain damage in a car accident, a musician became unable to hear a string quartet as anything but separate voices: "Four thin, sharp laser beams, beaming to four different directions." Perhaps the most distressing syndrome is 'dystimbria', in which musical tones are perceived as unpleasant noises, akin to banging pots and pans or the screeching of a car. Incredibly, Sacks experienced two episodes of temporary dystimbria, during which piano music sounded like "toneless banging with an unpleasant metallic reverberation".

Of course, music speaks to our emotions as well as our intellect, and Sacks presents evidence that the emotional response has its own network in the brain. He also reminds us that music engages our motor circuits (think of foot-tapping and dancing) — a phenomenon that underlies the use of music therapy to 'unlock' patients who have been immobilized by Parkinson's disease.

Sacks's great gift — as a physician as well as a writer — is to see the whole person, rather than just the brain or the disease. He certainly avoids the danger he identifies in his preface: that, as new technologies allow us to examine the brain in ever-greater detail, "the simple art of observation may be lost … clinical description may become perfunctory, and the richness of the human context ignored". Amidst the wealth of wonderful observations and the superabundance of humanity, another danger lurks: that the general reader who was hoping for a clear picture of what we know about the neuroscience of music may come away wanting more.

Help is at hand, in the shape of Levitin's very different but equally enlightening book. Levitin is ideally placed to write about music and the brain, having been a professional musician, sound engineer and record producer before embarking on a career as an academic neuroscientist. He starts with the basics: the physics and neuropsychology of pitch, timbre, rhythm, loudness and harmony. He then discusses the importance of processes such as perceptual completion ('filling in'), feature extraction and feature integration, and describes where they take place in the brain. The comprehensive bibliographic notes provide easy access to the relevant technical literature.

Levitin is just as interested as Sacks in the emotional power of music, but where Sacks tells us of unfortunate individuals who have lost the ability to be moved by music, Levitin offers a mechanistic explanation for its effects. "Music communicates to us emotionally," he says, "through systematic violations of expectations." He also presents evidence for the neural circuit that mediates the emotional response, which involves the cerebellum (the brain region that 'notices' deviations from the expected timing), the mesolimbic system (known to be involved in pleasure, reward and addiction), and the release of dopamine by the nucleus accumbens. A separate circuit involving the auditory cortex looks after our intellectual response, the perception of structure and form.

Can the emotional effect of music really be summed up as the violation of expectations? When I'm listening to something as gut-wrenching as the final movement of Tchaikovsky's sixth symphony, it certainly doesn't feel that way. But Levitin makes a persuasive case that memory, anticipation and surprise are fundamental to our enjoyment of music.

Levitin writes clearly and entertainingly, and has a real flair for analogy. His dual background adds to the fun. In one chapter we find him at lunch with Francis Crick, discussing the cerebellum; in the next, he's at dinner with Joni Mitchell, figuring out why her harmonies sound like no one else's (and why the bass player Jaco Pastorius was essential to her sound).

Both Levitin and Sacks are struck by the ubiquity of music across cultures and times, and by the observation that our brains seem to be, in Sacks's words, "exquisitely tuned for music". They vigorously rebut Steven Pinker's view of music as "useless", with no adaptive value, by asserting its importance in courtship, social cohesion, cognitive development and recall, and in simply making life more bearable. What tipped the balance for me, though, was learning from Sacks that, having no accurate timepieces available, Galileo timed his experiments on the descent of objects by humming tunes to himself. What could be more useful? ∎

Laura Garwin is a postgraduate trumpet student at the Royal College of Music in London. She is a former North American editor of *Nature* and former executive director of the Harvard Bauer Center for Genomics Research.

Using maths to tackle cancer

Dynamics of Cancer: Incidence, Inheritance, and Evolution
by Steven A. Frank
Princeton University Press: 2007. 400 pp.
$99.50, £59.95

Robert A. Weinberg

Multicellular animals have been around for about 600 million years, and cancer has been a problem for most of this time. There is a risk of cancer whenever the component cells grow and divide, so cancer incidence has increased progressively as the size and lifespan of organisms extended. In a lifetime, humans experience up to 10^{16} mitoses, the process by which a cell duplicates its DNA and divides to make two identical daughter cells. Each mitosis is an invitation to genetic disaster, due to miscopying of DNA, inadvertent breakage of chromosomes, and mishaps in chromosomal segregation. Given these vast opportunities for accumulating mutations, it is surprising that we don't generate numerous life-threatening cancerous cell clones during our first years of life.

The reason for our long and generally cancer-free lives is a series of anticancer defence mechanisms that have co-evolved with our increasing complexity. Most of these defences are wired into the intracellular signalling circuits that govern cell behaviour, although the organization of our tissues and immune systems contributes too. At least five or six of these mechanisms must be breached before a

question of whether we know enough biology and biochemistry to create truly useful, predictive models of multi-step tumour development. We have learned much over the past three decades about the biochemistry of cell transformation and the physiology of neoplastic cell growth, but we still have only a crude understanding of how these processes play out in living tissues. We don't understand in any quantitative way how endogenous mutagenic processes and environmental mutagens work. Nor do we understand the mechanisms or the kinetics governing the methylation of gene promoters and the resulting silencing of tumour-suppressor genes.

The fate of most of the pre-neoplastic cell clones remains an enigma. We don't understand how specific mutant alleles or mutant oncoproteins affect the darwinian processes of clonal selection and expansion. In fact, we don't really know how many distinct biological steps drive or accompany the formation of any type of human tumour. We know the approximate number of the slowest, rate-limiting steps, but almost nothing about the more rapid ones, yet these are just as important in driving multi-step tumour progression.
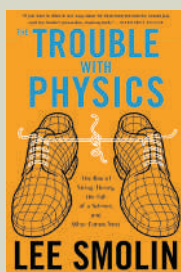
Frank is an evolutionary biologist with a strong mathematics bent. His book grapples with many of these issues, offering insights from his mathematical modelling of various steps of tumour progression. Those who aspire to digest all of his arguments, many of which are embedded in equations, will not find this easy reading. The text often appears in the formal voice of a mathematical proof.

A school of mathematical biology arose at the University of Chicago in the 1960s, only to decline in the following decade. It became apparent that a multitude of parameters needed to be assumed or arbitrarily fitted to existing data sets to ensure that the predictive powers of mathematical models conformed with actual observations. In the end, the proponents of mathematical biology drifted

full-blown
tumour appears.

These multiple lines of defence explain the complexity of tumour formation, which in most tissues involves a sequence of steps, each of which breaches one or more of the defences. We can describe some of these steps in biochemical detail, such as the activation of a *ras* oncogene or the inactivation of the *p53* tumour-suppressor gene. But a description of the process as a whole is currently well beyond our reach.

Steven Frank, in his new book *Dynamics of Cancer,* is the latest in a line of biologists and mathematicians who have taken on the challenge. Can we describe tumour formation in terms of rate equations and probabilities? Are

there aspects of each step that can be reduced to mathematical formulation? And can we develop a general model of multi-step tumorigenesis as it occurs in a variety of human tissues?

Solutions to these problems have ramifications beyond our understanding of tumorigenesis. For example, if we could assign values to the parameters governing the rates of each step of tumorigenesis, including the ways environmental and genetic factors affect these parameters, we could predict why cancer strikes at certain ages and in certain individuals.

Frank's approach to this problem raises the

"Can algebraic formulae tell us more than reasoning about the behaviour of complex biological systems?"

---

away, disenchanted.

Perhaps the biologists posed the most pointed question: can algebraic formulae tell us more than reasoning about the behaviour of complex biological systems? This question echoes almost half a century later, but Frank is undeterred. One day he may be seen as the pioneer who began the difficult task of building a sturdy foundation for a truly useful mathematical model of cancer development. ■

Robert A. Weinberg is at the Whitehead Institute for Biomedical Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA.

One-man canary

Suffer and Survive: The Extreme Life of J. S. Haldane
by Martin Goodman
Simon & Schuster: 2007. 320 pp. £14.99
(Not available in the United States)

Andy Meharg

That the nineteenth-century physician J. S. Haldane lived to his mid-70s was more down to luck than self-preservation. In his study of respiration and lung disease, Haldane performed series after series of experiments on himself, any one of which could have been fatal. In an air-tight chamber he systematically exposed himself to dangerous levels of toxic gases for extended periods to see how his body would react, his only precaution being a colleague watching through a window to see when he collapsed.
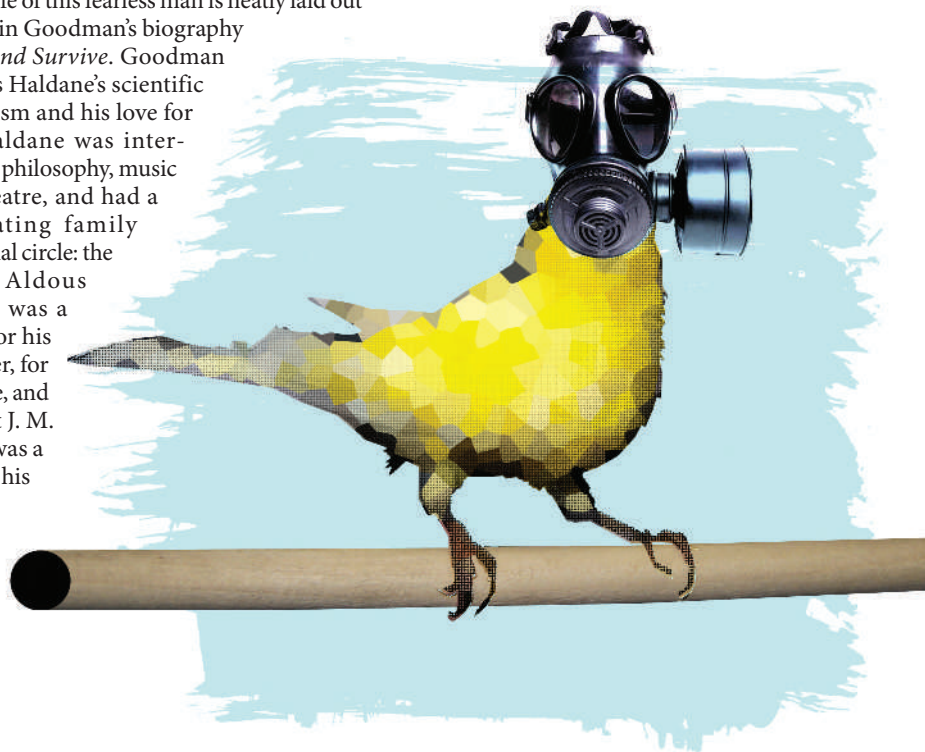
What drove Haldane was his innate scientific curiosity and a desire to save others' lives. He worked tirelessly to protect miners from the perils of coal dust, methane, carbon monoxide, carbon dioxide and low oxygen levels. He would rush to mine disasters to experience the fatal vapours and so determine the cause of death. He devised safety measures, such as his famed canary cage with a built-in oxygen supply to revive the bird when it collapsed. And by telling Cornish miners to spray the rock face with water as they cut, to reduce dust, he saved thousands from an early and painful death from the lung disease silicosis.

When German chemist Fritz Haber developed chemical warfare during the First World War, the British government turned to Haldane to protect the troops. There followed an arms race as Haber tried to outwit Haldane's development of gas masks. In the course of this work, Haldane frequently exposed himself to chlorine and mustard gases used on the battlefield. He conducted research on the front line to work out what gases were being used, and gauged their physiological impact by autopsying dead soldiers. He also pioneered the oxygen treatment used to recuperate soldiers who survived gas attacks.

Haldane's war effort went beyond the trenches. His work on deep-sea diving and the atmosphere in submarines saved many lives, and his studies of high altitudes benefited pilots and laid the way for the conquest of Mount Everest. It was Haldane who discovered that the body produces extra haemoglobin during acclimation to low oxygen at high altitudes.
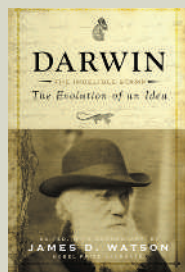
The life of this fearless man is neatly laid out in Martin Goodman's biography *Suffer and Survive*. Goodman conveys Haldane's scientific dynamism and his love for life. Haldane was interested in philosophy, music and theatre, and had a fascinating family and social circle: the writer Aldous Huxley was a suitor for his daughter, for example, and novelist J. M. Barrie was a guest at his Scottish estate. In 1912, using a scientific mission as cover, Haldane accompanied his brother Richard — secretary of state for war and later the first Labour lord chancellor — to Germany for secret negotiations with the Kaiser to avert war.

This fine book is not without fault. The first chapter has vividly imagined scenes embroidered onto real events, presumably for dramatic effect, describing how Haldane rushed to a Welsh coal-mining accident. This device is thankfully used less in subsequent chapters. However, it is not until 100 pages in that descriptions of Haldane's experimental gas chamber and his measurements of carbon dioxide and carbon monoxide in blood, so central to the opening chapter, are outlined. After this shaky start, the book becomes highly enjoyable. It is a fitting tribute to a pioneer who enabled the human body to survive at the extremes of modern life. ■

Andy Meharg is professor of biogeochemistry at the University of Aberdeen, Cruickshank Building, St Machar Drive, Aberdeen AB24 3UU, UK, and is the author of *Venomous Earth: How Arsenic Caused the World's Worst Mass Poisoning*.

Darwin: The Indelible Stamp
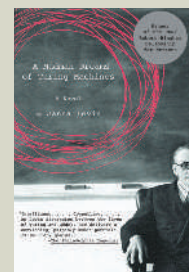edited by James D. Watson (Running Press, $21.95, £11.99)
With this collection of four of Darwin's great works, the editor, James Watson, sought "to provide some historical context and connection to current science, as well as some personal observations about how Darwin's writings affected their intellectual development" (Bruce H. Weber, *Nature* **438**, 287; 2005).

A Madman Dreams of Turing Machines
by Janna Levin (Anchor, $13.95)
It's a tale of two thinkers, as cosmologist Janna Levin fictionalizes the lives and ideas of Kurt Gödel and Alan Turing. The great logician and the extraordinary code-breaker never met, but in Levin's poetic novel their stories are imaginatively intertwined. Explaining her switch from popular science, the author has said, "I didn't believe a reader would get the same visceral impact from a pedagogical lesson."

Botanists' blues

Nature's Palette: The Science of Plant Color
by David Lee
University of Chicago Press: 2007. 384 pp. $35

Philip Ball

"Why grass is green, or why our blood is red, Are mysteries which none have reach'd into." John Donne's words were true in the seventeenth century. Today they certainly aren't, as David Lee makes clear in *Nature's Palette*, an enchanting survey of colour in plants. Among the mysteries is why some plants are electric blue, such as *Selaginella*, a relative of ferns that thrives in the rainforests of Malaysia. Its leaves are as blue as some butterfly wings, and for the same reason: it is a 'structural' colour, produced by light scattered from a periodic banded microstructure in the outer cell walls. The shaded undergrowth of the forests is home to several other blue-leafed plants, but it's not clear why they should reflect so much light when it's such a scarce commodity. After 30 years of looking, Lee still has no good explanation.

Blue leaves abound in seventeenth-century Dutch flower and landscape paintings. One of the most striking examples is Adam Pynacker's *Landscape with Sportsmen and Game*, in which a garish clump of blue foliage on the gloomy woodland floor appears every bit as surreal as *Selaginella*. This is not a botanical oddity, however, nor is it what the artist intended. Over time the yellow pigment in the green mixture has faded. But there is a botanical explanation, as painters commonly used yellow pigments from plant extracts, such as saffron or weld, that are damaged by light. Plants require merely a seasonal burst of exuberant advertising, so their pigments don't last — as generations of artists and dyers have found to their cost.

Dutch flower paintings also depict tulips streaked with feathery white. These 'broken tulips' appear unpredictably among otherwise identical bulbs, and are short-lived and hard to propagate. As a result, they were so rare in Baroque Holland that it was often cheaper to hire a painter to capture their beauty than to acquire a bulb. Speculators bought thousands of bulbs in the hope of finding a few such variants that would sell for a fortune, creating the infamous 'tulip mania', the archetypal economic bubble driven by speculation. As Lee explains, we know now that the broken tulips were diseased, infected by a virus that interferes with pigment production.

Lee's book is packed with many such gems from botanical and social history. So captivating is his passion for botany that his occasionally bewildering thickets of carotenes and anthocyanins can be forgiven. His paean provides a compelling case that botany is full of intellectual challenges, many shamefully neglected. He shows, for example, that plants are extraordinary biological devices, with leaves that act as arrays of lenses or mobile light antennae, or that change colour on exposure to light, thanks to the mobility of their chloroplasts. But only one Nobel prize has ever been awarded for fundamental biological research on plants — for Barbara McClintock's discovery of transposons in maize, which had to weather a storm of scepticism before being rewarded.

Like its subject, the book can be overwhelmingly heady. A scary glimpse into the hothouse world of Miami orchid breeding suggests there is a lurid thriller waiting to be written on the subject by Carl Hiaasen. And the cloying scents of the 'gallant south', evoked so unforgettably in the Billie Holiday song *Strange Fruit*, are all the more palpable for knowing that some flower odours mimic rotting carcasses to attract carrion flies. So obsessive is the flower-lover's demand for novelty that a biotech company, Florigene, set itself up with the sole mission of making blue roses and carnations. A jackpot awaits if they succeed, but I can't help thinking the romance will be well and truly lost on the way. ■

Philip Ball is a consultant editor of *Nature*.

What do mathematicians do?

How Mathematicians Think: Using Ambiguity, Contradiction, and Paradox to Create Mathematics
by William Byers
Princeton University Press: 2007. 425 pp. $35

The Mathematician's Brain: A Personal Tour Through the Essentials of Mathematics and Some of the Great Minds Behind Them
by David Ruelle
Princeton University Press: 2007. 176 pp. $22.95

The Mind of the Mathematician
by Michael Fitzgerald & Ioan James
Johns Hopkins University Press: 2007. 312 pp. $30

Donal O'Shea

Mathematics is a corpus of immutable, rigorously established facts, right? Wrong. Three recent books argue that mathematics is a profoundly human enterprise, best described as what mathematicians do.

In the most ambitious, accessible and provocative of the three, *How Mathematicians Think*, William Byers argues that the core ingredients of mathematics are not numbers, structure, patterns or proofs, but ideas. Numbers are ideas, and so are functions and even logic. Ideas are organizing principles. They are neither right nor wrong.
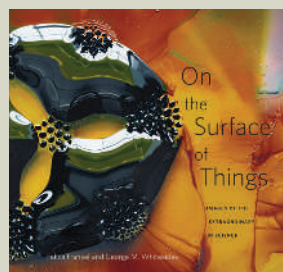
The polymath Arthur Koestler famously characterized ambiguity as involving a single situation or idea that is perceived in two self-consistent but mutually incompatible frames of reference. Mathematics thrives on ideas that harness ambiguity. Byers maintains that great ideas, like Simone Weil's notion of 'true good', have conditions that are contradictory and so seem impossible — they encapsulate ambiguity that is so great as to seem paradoxical.

There's ambiguity everywhere. Zero, for instance, captures both the notion of nothingness, or absence, and the idea that nothing is something. 89/17 signifies both a number and the operation of dividing 89 by 17. Algebraic expressions such as $3x+2$ represent a process (take a number, multiply it by 3 and add 2) and a set of numbers. The variable x in an

On the Surface of Things
by Felice Frankel and George M. Whitesides (Harvard Univ. Press, $24.95, £16.95)
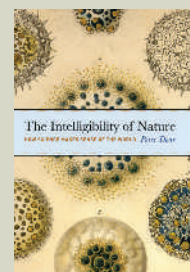A handsome re-issue of Felice Frankel's startling photography reveals the extraordinary beauty of chemical surfaces. Roald Hoffmann wrote: "One holds in one's hands an obvious classic at the nexus of art and science" (*Nature* **389,** 348; 1997).

The Intelligibility of Nature: How Science Makes Sense of The World
by Peter Dear (Univ. Chicago Press, $17)
Science historian Peter Dear shows how the twin objectives of 'doing' and 'knowing' gave rise to modern science as a means of understanding the world. "Scientists who wish to reflect on their vocation will gain valuable insights from this beautifully contrived book" (Richard Yeo, *Nature* **445,** 820–821; 2007).

equation simultaneously represents every number and a specific one, and the equality sign also embodies ambiguity. Ideas such as infinity, limits, derivatives and real numbers mediate between seemingly incompatible frames of reference. Their inherent ambiguities require a creative effort to master but give the ideas their power.

Great ideas often emerge in times of intellectual crisis, when ambiguity and apparent impossibility challenge the human impulse to understanding. In Byers' view, the formalist programme to reduce mathematics to formal axioms and their consequences was a great idea that responded to the seeming impossibility of non-euclidean geometries. But neither formalism nor any other great idea comes close to exhausting mathematics. "The content of mathematics," writes Byers, "cannot be definitively separated from how mathematics is created and understood."

Byers' view springs from the various facets of his career as a researcher and administrator (and, he says, his interest in Zen Buddhism). But it is his experience as a teacher that gives the book some of its extraordinary salience and authority. He points out that the concepts that cause students trouble are often those that took a long time to develop historically. Good mathematics teaching should not banish ambiguity, but enable students to master it. Conflating mathematics with its logical structure results in teaching that values the presentation of logical structure over understanding.

The Mathematician's Brain by David Ruelle tackles some of the same questions as Byers' book but has a different emphasis. Ruelle imagines an enormous infinite-dimensional space of all formal axiom systems and all possible formal consequences of them. The human brain, unlike a computer, is not well adapted for working directly with formal statements. Human mathematics is a discussion in natural language, as used in everyday communication, about a formalized text, which remains unwritten (but which could in principle be written). Mathematicians work primarily with ideas, which Ruelle defines as short statements in human mathematical language that can be used in a human mathematical proof. Human

mathematics is a labyrinth of such ideas that allows mathematicians to move within the huge space of formal statements. Ruelle notes the extraordinary amount of context that underwrites mathematical activity, and reminds us that there is no reason to expect that a short human mathematical statement admits a short proof. The text is enlivened by many unusual mathematical examples, and by Ruelle's reflections on his own and other famous mathematicians' experiences.

If mathematics is what mathematicians do, are there any psychological traits or personalities that characterize mathematicians? Ruelle addresses this lightly with some illuminating insights, whereas Byers does not. *The Mind of the Mathematician* by eminent topologist

Ioan James and psychiatrist Michael Fitzgerald focuses almost exclusively on this question. The book begins with a systematic survey of the literature on the psychology of mathematicians. It then surveys the literature on mathematical education and considers the psychodynamics of mathematical creativity. In both instances, the authors focus on the exceptionally able and pay attention to issues of gender. The second half of the book consists of exquisite biographical sketches of 20 prominent mathematicians. The authors' careful treatments are an especially welcome addition to a genre riddled with apocryphal anecdotes and shoddy scholarship.

All three books are useful additions to the literature and have little overlap. Those with a serious interest in the psychology of creativity should read the book by James and Fitzgerald. Mathematicians and theoretical physicists will enjoy Ruelle. Everyone should read Byers. You may not agree with him, but his lively and important book establishes a framework and vocabulary to discuss doing, learning and teaching mathematics, and why it matters. ∎

Donal O'Shea is dean of faculty and mathematics professor at Mount Holyoke College, South Hadley, Massachusetts 01075, USA. He is the author of *The Poincaré Conjecture: In Search of the Shape of the Universe*.

Physical interactions

Structures of Scientific Collaboration
by Wesley Shrum, Joel Genuth & Ivan Chompalov
MIT Press: 2007. 296 pp. $35, £22.95

David Harris

As scientists across the world work on ever more ambitious projects, they are collaborating more and on larger scales. How does that collaboration emerge and develop? And how does the nature of the collaboration
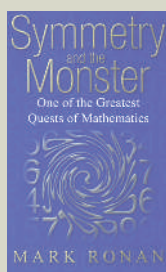
affect the success of a project?

Scientists often have a narrow view of collaborations, based on their own experiences. Most studies of collaborations have also focused on individual cases. *Structures of Scientific Collaboration* by Wesley Shrum, Joel Genuth and Ivan Chompalov surveys 53 collaborative projects and reveals remarkable diversity. The authors analysed oral histories collected in the 1990s for the American Institute of Physics and covering a range of fields including

---

Symmetry and the Monster: One of the Greatest Quests of Mathematics
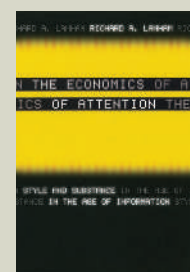by Mark Ronan (Oxford Univ. Press, $19.95, £8.99)
The quest to understand the mathematics of symmetry began 200 years ago in revolutionary France. Mathematician Mark Ronan tells the story of how a number of coincidences and a huge international collaboration finally led to the discovery of the 'monster' of symmetry — a giant snowflake with 196,884 dimensions.

The Economics of Attention
by Richard A. Lanham (Univ. Chicago Press, $18)
This interdisciplinary book from a professor of English attempts to chart the rise of intellectual property, arguing that, in this age of information, the one resource we are short of is attention — the ability to focus on what's important. An increasing importance of style over stuff will see the arts and letters grow to be more crucial than engineering and the physical sciences. It is quite a claim.
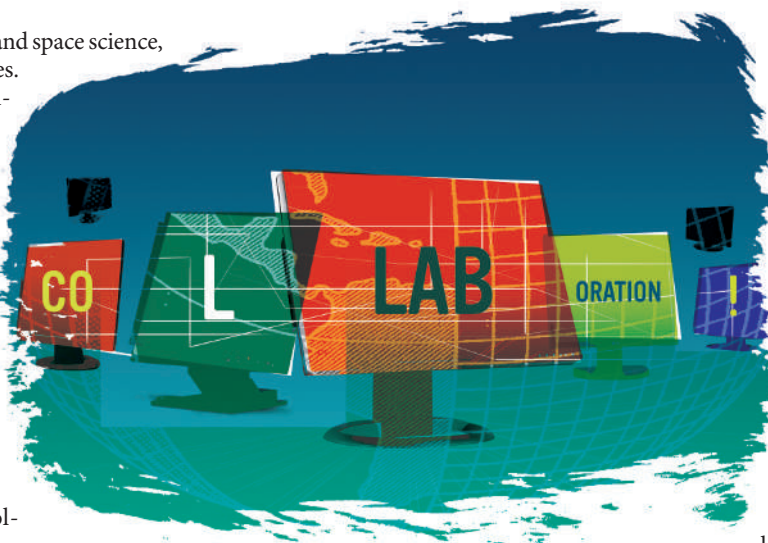
geophysics, particle physics and space science, primarily in the United States.

They developed a quantitative approach using statistical and organizational concepts. They shy away from extended case studies, as it can be difficult to generalize from them, but their studies retain some individual character, to give far more than a dry bibliometric analysis of who did what with whom. This approach helps them provide an enlightening, easy to read and sometimes surprising view of how collaborations work.

Trust, conflict and performance are crucial factors. One counter-intuitive finding is that trust does not correlate with performance. It is widely assumed that trust is required to achieve success, but this may stem from the fact that lower levels of trust tend to result in higher levels of conflict. Conflict is unpleasant for individual scientists but is not necessarily a barrier to progress. In fact, many collaborations inherently tend to protect against damage from conflict between members.

There are several types of organizational structure, characterized mostly by the breadth of activities and manner of governance. Most types occur in a range of scientific fields. One sort — the 'quasi-Athenian democracy', typified by bottom-up consensus building — is almost unique to particle physics. The book's authors argue that this model would not benefit other science communities, and indeed, as particle-physics collaborations grow larger, they might need to change from this traditional form.

One weakness of this characterization of particle physics is that collaborations have already grown in size well beyond the cases studied, all of which occurred in the 1970s and 1980s. It would be interesting to compare two competing but nearly identical projects, such as the BaBar experiment at the Stanford Linear Accelerator Center in California and the Belle experiment at the High Energy Accelerator Research Organization (KEK) in Tsukuba, Japan. These have different organizational and management styles, with different strengths and weaknesses.

The study's applicability is limited by the US-centric and physics-dominated data set. It would have been interesting to read about international collaborations, such as the ITER fusion-reactor project in Cadarache, France, the Human Genome Project and the Large Hadron Collider at CERN, the European particle-physics laboratory near Geneva. Do those projects fit the authors' models even though international diplomatic and political factors have played a more prominent role?

Under what conditions is collaboration desirable? This, too, needs further exploration, especially as more projects are located and managed in large laboratories. A recent report on university participation in US particle physics by the Department of Energy's High Energy Physics Advisory Panel suggests that the concentration of collaborative projects in large laboratories could be weakening the physics effort in universities and contributing to structural deficiencies in the science programme. ■

David Harris is editor-in-chief of *Symmetry* magazine at the Stanford Linear Accelerator Center, 2575 Sand Hill Road, Menlo Park, California 94025, USA.

Eras of judgement

Objectivity
by Lorraine Daston & Peter Galison
Zone Books, MIT Press: 2007. 500 pp.
£25.95, $38.95

Theodore M. Porter

All scientists think they know what objectivity is. But objectivity has a history full of fascinating changes of sense, and now bears several different meanings. Most involve an absence: of distortion; of human bias or subjectivity; or of the particularities of location and culture. It is one of the more austere virtues, often seen as a stern judge that disallows ornamentation.

In their book *Objectivity*, Lorraine Daston and Peter Galison pursue the idea that we can best explore the many meanings of objectivity through an examination of images — their book is essentially about the ideal of objectivity as applied to scientific atlases, or reference books of illustrations. In such works, objectivity is largely about restrictions on the composition and presentation of scientific images. *Objectivity*, however, offers a cornucopia of images ranging from plants and birds to embryos and snowflakes, and it is enriched by the authors' expert sleuthing and wide intellectual reach.
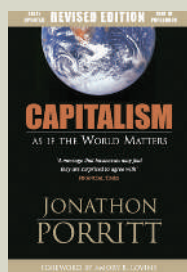
The heart of objectivity, for Daston and Galison, is 'mechanical objectivity', which had its heyday from about 1850 to 1920. The phrase implies that scientists proceeded methodically, using automation whenever possible, to reduce human intervention to a minimum. This era of impersonal objectivity was preceded, the authors propose, by one of 'truth to nature' and succeeded by one of 'trained judgement'.

Using such names for historical periods may seem incendiary, first by implying that objectivity has had its day, and further by placing it in opposition to both truth and expert judgement. But the authors have very particular meanings in mind. The terminology is confusing, though, because scientists rarely used the word 'objectivity' until some decades into the

985

era that Daston and Galison name for it.

Objectivity here is more about morality than validity. Mechanical objectivity made suppression of self a duty, a triumph of abstemious will over the prideful claim of competence to judge and interpret. The contrast between mechanical objectivity and devotion to truth is subtle but important. A uniformly sharp photograph might well be confusing, masking the main point in irrelevant detail. It is easier to identify a flower or recognize a diseased kidney from an idealized drawing than from an unretouched photograph.

Daston and Galison define the era of mechanical objectivity by a profound reluctance of scientists seeking to illustrate their work to stray from the rigorous accuracy of the camera. Yet photographic accuracy, at least until 1890, was far from rigorous. Choices had to be made in preparing materials, composing, illuminating and framing the picture, and sometimes retouching an image, and the technology of photographic reproduction in the 1880s involved intense human labour. The language of objectivity, then, did not so much distinguish photography from drawing as mark off certain ways of preparing photographs from others. An 'objective' image was one that minimized the effects of human intervention.

In contrast, naturalists of Daston and Galison's earlier 'truth-to-nature' era felt no such compunctions. They used illustrations to get at something more fundamental than the individual: an underlying type. This is 'truth' in a very special sense. The eighteenth-century naturalists Carl Linnaeus and René Réaumur supervised their artists meticulously, not to ensure that they drew a specimen accurately, but to make sure that they departed from its particularity where this could reveal a deeper level of truth.

The authors argue that science and art diverged around 1850, with science insisting on objectivity in its ascetic form, and art rejecting the mere reproduction of what can be seen in the world. They propose that different forms of objectivity arose in conjunction with new understandings of the 'scientific self', which by 1850 had become quite different from the artist. This focus on the changing character of the scientist offers important insights, but the authors overstate their case for the rupture between science and art. The very phrase 'truth

to nature' is associated specifically with the pre-Raphaelites, a mid-nineteenth-century artistic movement grounded in an admiration for the writer and aesthete John Ruskin (indeed Tate Britain had a major exhibition in 2004 called "Pre-Raphaelite Vision: Truth to Nature"). Ruskin, whom Daston and Galison fail to mention, was as committed to the union of scientific and artistic perspectives as the poet-naturalist Johann Wolfgang von Goethe, who features prominently.

How could objectivity interfere with truth? Daston and Galison introduce this notion at the outset with the story of the British physicist Arthur Worthington. In 1875, Worthington began chronicling the beautiful symmetries of fluid motion revealed by a droplet of mercury hitting a clear glass plate. In 1894 he finally captured the splash in a photograph. To his dismay, it was a mess, with none of the perfection he thought he had witnessed. Did the mechanical image occasion a sacrifice of truth? Yes, if 'truth' refers to such elegant forms as admired by Goethe a century earlier. In a more prosaic sense, Worthington had finally got it right. He concluded, sadly, that the symmetries he thought he saw were artefacts of his own flawed vision.

The authors draw back from claiming in a strict sense that objectivity was invented, had its day in the sun, and then passed away. Yet something like this, much qualified, is the organizing principle of the book. This is problematic, not least because the argument is based on the special case of scientific atlases and then extended to all of science. Can Daston and Galison really claim that trained judgement never existed as a conscious alternative to unthinking routines of objectivity until about 1920? They are, however, persuasive in arguing that a vast expansion of scientific education in the early twentieth century helped to make
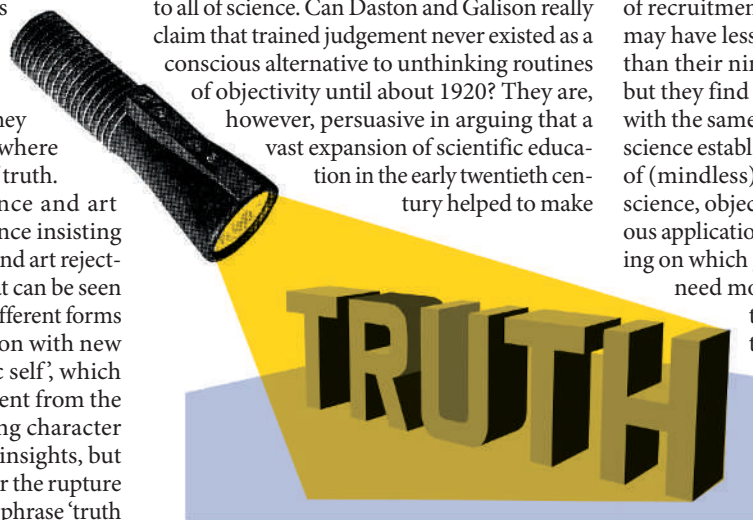
expert judgement more acceptable to science.

By this time, scientists and other professionals were arguing with increasing conviction that mechanical procedures could not really get at what is most interesting. The interpretation of images, they said, depended on a capacity to see 'family resemblance', which scientists, with the aid of well-drawn images, could learn to recognize reliably. Daston and Galison demonstrate, with wonderful insight, that this notion of family resemblance was simultaneously scientific, philosophical and pictorial. Ludwig Wittgenstein, in enunciating it, referred back to some striking composite photographs through which the statistician and eugenicist Francis Galton tried to establish racial and social types. Scientists of this period repeatedly likened the analysis of an image in physics or medicine to detecting racial background from a photograph.

The book concludes with an analysis of images from today, perhaps a fourth epoch in the history of objectivity. Modern techniques of image production at the nanoscale use processes that record themselves (Daston and Galison call these 'presentations' rather than 'representations'). These attest to the extraordinary merging, in certain areas, of technology with science, and sometimes also with marketing. The commercialization of science presents its own problems, for scientific objectivity is about more than communication among scientists. As important today are the views of political leaders and the public. Modern scientists, enmeshed in elaborate systems of recruitment, training and communication, may have less need of mechanical objectivity than their nineteenth-century counterparts, but they find it difficult to speak to the public with the same kind of assurance. Critics of the science establishment have turned the rhetoric of (mindless) mechanical objectivity against science, objecting to the sensible but unrigorous application of experience and understanding on which good science depends. Scientists need more than ever before to gain, and to deserve, public respect for their skilled work and uncorrupted expert judgement. ■

Theodore M. Porter is in the Department of History, University of California, Los Angeles, Los Angeles, California 90095-1473, USA.

Things That Talk: Object Lessons from Art and Science
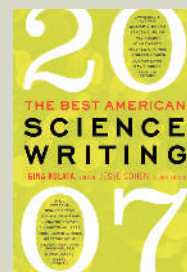edited by Lorraine Daston (Zone Books, $21.95)
This collection of nine essays, edited by science historian Lorraine Daston, explores how objects attract meanings and cultural significance — and shows how they then become crystallization points for discussion. Each essay singles out one thing for close attention, including a Bosch drawing, soap bubbles, Rorschach blots and paintings by Jackson Pollock.

The Best American Science Writing 2007
edited by Gina Kolata (Harper Perennial, $14.95)
The fifth anthology of features and essays about science is compiled by *The New York Times*' science writer, Gina Kolata. *The New Yorker* wins six of the 20 slots, including one for its unforgettable article 'Manifold destiny' about Grigory Perelman and the Poincaré conjecture. Also prominent is writing about the brain, notably by Joshua Davis of *Wired* on the condition of face blindness.

The power of the mind
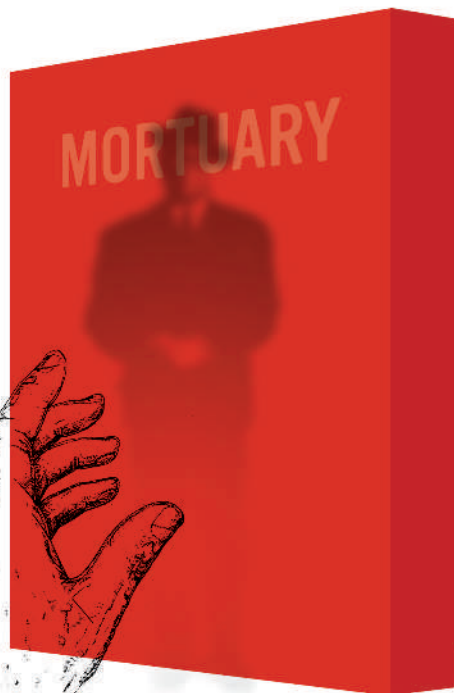
Ghost: A Novel
by Alan Lightman
Pantheon: 2007. 256 pp. $23

Janna Levin

Few writers are able to harness fully the narrative potential of science in fiction. Some stunning novels touch on science tangentially, investing more in the grim consequences, such as Cormac McCarthy's post-apocalyptic masterpiece, *The Road*, and Kazuo Ishiguro's haunting twist on genetic ethics, *Never Let Me Go*. As a gifted physicist and writer, Alan Lightman brings unique sensibilities to this context, structuring his works around scientific themes. His visionary first work of fiction, published in 1993, was *Einstein's Dreams*, a series of poetic vignettes, meditations on the theory of relativity. The book's success lies in Lightman's ability to draw out the fantastical aspects of scientific theory while simultaneously reflecting some simple facet of our intimate human desires — an accomplishment that won him many accolades. In his latest novel, *Ghost*, Lightman considers another aspect to the overlap of sense and sensation, a more troubled but delicately illustrated tension between reason and belief, between the ordinary and the extraordinary.

Ghost is the story of an extraordinary event experienced by an ordinary man. David has seen something he cannot name or explain. He is not given to fantasies of the supernatural. He is a reasonable person, intelligent, hard working and fair-minded, if lacking in ambition. Languishing after a painful divorce and unexpectedly losing his long-held job at the bank, he stumbles into an apprentice position at a local mortuary. And it is there, where the bodies rest for viewing, that he sees something inexplicable.

He is terrified, not by the vision but by the threat to his rationality. He is analytically minded and entirely capable of critical thinking, but he is sure that it was no trick of the light or hallucination. Deeply embarrassed about his experience, he hesitantly describes the vision to Ellen, the woman with whom he has an emotionally sparse but highly sexual relationship. The story quickly spreads through the community of stragglers in the decrepit apartment building in which he lives alone in one shabby room. The local newspaper carries the headline: "Mortuary worker claims to see ghost".

David becomes an ambivalent ally to the Society of the Second World, while maintaining a confused yet sincere alliance with their opposition, a group of scientists at the local university. David agrees to let them carry out public experiments to measure his psychic ability to affect a random-number generator enclosed in a black box.

This is one interpretation of the story. There is another. David, although not very ambitious, is not without vanity. For years he exaggerates to his mother his importance at the bank, pretending that he is the manager. He struggles daily with the humiliation of his wife's rejection and seizes on the opportunity to appear important to her, to seem valuable and significant. He is not rabidly egotistical, just vulnerable enough to believe that he can affect a black box with his mind. He needs to believe that he is special — "You've been chosen," Ellen tells him — and so he does believe it.

His failure to fully understand the scientists' statistical arguments allows him to indulge in this fairly mild delusion of self-importance. He is genuinely uncomfortable with the public attention he receives, but needs to be believed by those close to him. Even here he lacks ambition. He has no inclination to claim that he has psychically altered major world events, as the Society of the Second World's propaganda declares, just that he can affect the black box. He cannot quite accept that his experience was nothing significant.

Once again, Lightman cleverly navigates a precarious line between science and belief. The facts of David's experience are never resolved. Lightman allows just enough ambiguity for you to impose your own beliefs, in the absence of proof. He could easily mock his characters' weaknesses, but instead he evokes our sympathies and challenges us all, rationalists included.

Lightman's writing has an elegant and light touch. There are lovely descriptions of the processes the corpses undergo: the embalming, the sewing together of the lips, the colouring of the faces. All the while, in and out of the mortuary, there are subtle reflections on death and our fears about life. As we are lost in this ordinary man's world, the story sneaks up on us, understated and beautiful, until, before we realize it, something extraordinary has happened.

> "Lightman considers a tension between reason and belief, between the ordinary and the extraordinary."

Janna Levin is a professor in the Department of Physics and Astronomy at Barnard College of Columbia University, 3009 Broadway, New York 10027, USA, and is the author of the novel *A Madman Dreams of Turing Machines* (Knopf, 2006).

The Cutter Incident: How America's First Polio Vaccine Led to the Growing Vaccine Crisis
by Paul Offit (Yale University Press, $17)
In 1955, shortly after the development of the polio vaccine, some 200,000 people were mistakenly injected with live, virulent polio vaccine. Vaccine researcher Paul Offit discusses what went wrong, and shows how it has had a devastating effect on vaccine production ever since. "A superb job," wrote Paul-Henri Lambert (*Nature Med.* **12,** 879; 2006).

Every Second Counts: The Extraordinary Race to Transplant the First Human Heart
by Donald McRae (Pocket Books, £7.99)
Forty years ago the race was on to perform the first human heart transplant. *Every Second Counts* is a gripping account of how the South African doctor Christiaan Barnard came from nowhere to secure his place in medical history. McRae details how Barnard pipped to the post three other surgeons who were also poised, having spent years perfecting their techniques.

Fishing for certainty

Science advisers should have confidence in their data, or risk being undermined by more dogmatic and vociferous stakeholders during the policy-making process.

Andrew A. Rosenberg

Policy-makers receive formal and informal advice from all quarters: scientific, legal, political and public. Each piece of advice is considered mandatory by the giver, and it often conflicts with other advisers' points of view. Uncertainty is a feature of all advice, but it is usually only acknowledged by the scientific adviser.

I have worked as a scientist, policy-maker and adviser, mostly managing marine resources. As an ecologist specializing in fisheries population dynamics, I naively assumed that scientists develop advice that is passed on to policy-makers who then make decisions in the light of it.

When in 1995 I moved into policy-making, managing fisheries in the northeastern United States, I learned that advice comes from all directions. Scientists would present data with many caveats; others would give advice based mainly on opinion. Fishermen coming to the microphone in a public meeting might categorically state that the science was wrong, the rules wouldn't work and everyone would go out of business. Scientists tended to emphasize their uncertainty, and would be unwilling to speculate.

As scientists, we learn to analyse uncertainty and we explore decision-making in the light of that uncertainty. This is important, but we must also recognize that the precautionary approach will be adopted only slowly in policy-making. Uncertainty undermines political will in environmental decision-making. Officials are more likely to support a vociferous interest group that is apparently certain of the dire economic consequences of new restrictions, than scientists who advocate caution and prioritize the environment.

Over time, I learned that the solution for an adviser is not to hide careful analyses of uncertainty, but to distinguish the almost certain from the less certain. For example, it became clear in the 1980s that overfishing in New England, the North Sea and many other areas was critically depleting

D. PARKINS

resources. Exploitation of species such as cod was removing 60–70% of the standing stock every year. Unfortunately, the debates were too often about whether the sustainable exploitation rate should be 20 or 25%. The conclusion drawn by many in industry and politics was that the science was uncertain. Hearing people say in debates, "fisheries science is not an exact science," made me wonder which other field they were comparing fisheries to, and indeed what an exact science is.

There is little uncertainty that over-

> "Emphasizing what we don't know often drowns out what we do know."

fishing was, and in many cases still is, occurring and that exploitation needed to be reduced by half or more. Emphasizing what we don't know often drowns out what we do know. In the event, strong action in New England reduced exploitation rates on some stocks, such as haddock, down to reasonable levels. As scientists predicted, the stocks began to recover. On other stocks such as cod, exploitation has remained relatively high, and their numbers have not recovered. There is little mystery, and only very slow progress is being made. Unfortunately, the fish may not wait for us to learn our lesson.

Statements of policy are still a far cry from implementing policy. It is easier to agree to the general principle of ending overfishing and rebuilding resources than it is to put the principle into effect. Few argue that overfishing and resource depletion are good things; many argue about whether their fishing activity, their business or their recreation really contributes to overfishing.

For example, the US Marine Mammal Protection Act of 1972 is a strong mandate to protect all marine mammals; its reauthorization in 1994 was passed unanimously by the Senate. But in the northeastern United States, protection of whales from entanglement in fishing gear — one of the main causes of death for whales in coastal waters — means that restrictions on fishing are necessary. Implementing these restrictions caused huge controversy. Disagreement between different interest groups was exemplified by the elected official who opposed the restriction, telling me to, "go save the whales somewhere else".

Political decision-making inevitably leans towards minimizing the impacts of policies on those constituents who are most affected. The public cares about the general outcome, such as saving whales, but individuls are unlikely to change their political view or support for an official because of local issues such as catch quotas or protected areas; fishermen will because the issue is immediate and vital to them.

In the 1990s, when I was a senior manager of the US National Marine Fisheries Service, I viewed my job as maximizing conservation without someone higher in the policy-making structure taking away my authority. Each decision was a judgement call about how far I could go, and without a doubt my judgment was imperfect. Science led my logic. I would start by asking: what do we know, and what does that mean we should do? In every case, I would then have to consider: what can be done, given the forces at play? As an adviser, I learned that adhering closely to the scientific advice is always the best course — as long as you can save some fish in the process. ■

Andrew A. Rosenberg is professor of natural resources at the Institute for the Study of Earth, Oceans and Space, Morse Hall 142, Durham, New Hampshire 03824, USA.

SCIENCE & POLITICS

NEWS & VIEWS

PLANT BIOLOGY

Plumbing the pattern of roots

Bruce Veit

The results of a powerful combination of computer modelling and experimental tests can account for the establishment of gradients of the plant molecule auxin and for major patterning elements in the plant root.

Morphogens are molecules that govern pattern formation in organisms. They are perhaps best known through the textbook models of morphogen action that can generate a pattern as simple as the tripartite French flag or as complex as a segmented fruitfly embryo[1]. In real flies, the gene-transcription factor Bicoid acts as a morphogen. It forms a concentration gradient across the developing embryo through diffusion from a localized site of synthesis, then instructs distinct programmes of development as a function of its concentration. On pages 1008 and 1053 of this issue, Grieneisen *et al.*[2] and Galinha *et al.*[3] describe how they have applied an ensemble of molecular and modelling approaches to explore a different mode of gradient formation in plant roots. It relies on the directional transport of auxin, more familiar for its action as a hormone. This directional transport may thus enable auxin to act as a morphogen.

A wealth of physiological studies has highlighted auxin's concentration-dependent effects on the division, expansion and functional fate of cells, as well as its mobility within the plant[4–6]. In the root, there is a concentration maximum that coincides with the quiescent centre (QC), a small group of cells that sits just under the root cap and that can be thought of as an organizer for the surrounding array of initial dividing cells and their more differentiated products. This organization is sensitive to chemical treatments that perturb endogenous auxin levels, as well as to mutations in genes that are thought to execute specific sets of auxin-dependent responses[7].

Genes that mediate the directional movement of auxin also have a major influence, with genes encoding the PIN family of proteins having an especially prominent role. These proteins facilitate the movement of negatively ionized forms of auxin across the cell membrane, and hence promote its efflux from the cell. Once in the more acidic environment of the cell wall, auxin becomes protonated, and in an uncharged form can move across the membrane into adjacent cells via passive diffusion or auxin influx transporters[8]. The highly asymmetric distribution of PIN proteins in plant cells has served as the basis for
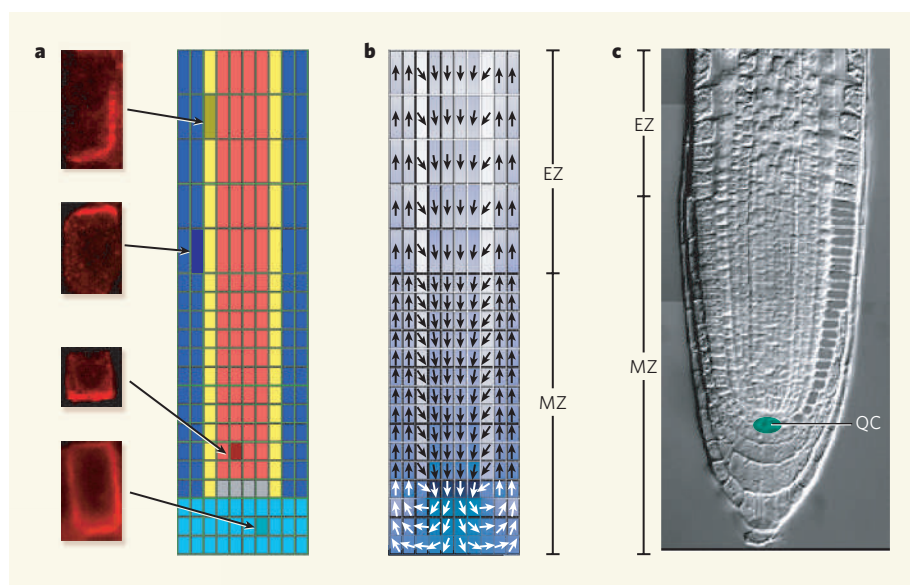

Figure 1 | The PIN-dependent formation of an auxin gradient, and patterning of plant roots. a, PIN proteins control auxin efflux from cells, and they are asymmetrically localized (fluorescence in images on left) in cells of different tissues in the root depending on how central or peripheral the tissues are. **b**, PIN asymmetry results in polar flows of auxin and the formation of auxin gradients; higher relative levels of auxin are indicated by darker background shading. **c**, The auxin gradient acts in some way to maintain directly proportional levels of transcription factors of the PLETHORA group of proteins, which act to establish and maintain distinct regions of the root. The highest levels of auxin determine the quiescent centre (QC), which is characterized by pluripotent cells that divide far less frequently than surrounding cells. Lower levels determine the meristematic zone (MZ), where cells divide more frequently. Still lower levels determine the elongation zone (EZ), where the elongation and terminal differentiation of cells occurs. (Figure adapted from Figs 1 and 2 of ref. 2, in which these events are modelled.)

'inverted fountain' models of auxin flow, which are in rough agreement with experimentally observed fluxes[9] (Fig. 1). In this model, basally positioned PIN proteins in cells of the central vascular elements mediate a downward flow of auxin, whereas apical PIN localization in more superficial epidermal and cortical layers promotes an upward movement. Finally, through laterally positioned PINs in cortical and sub-epidermal layers, there is an inward reflux, or reverse flow, of auxin into the downward flows in the vascular elements.

To explore just how well such a scheme might work and how it would affect auxin concentrations, Grieneisen *et al.*[2] developed a simple model that predicts auxin distributions as an outcome of directional auxin flows coded

at a cellular and subcellular level. Not surprisingly, like a crush of train commuters converging on the platform exit, an auxin maximum is predicted to occur where lateral and downward flows meet at the QC, before diverging to begin their upward journey. The real power of the model, however, lies with its demonstration of how robust this efflux-driven maximum can be. It can withstand a range of values for cell permeability and diffusion; floods of externally supplied auxins; and even cutting off the main source of auxin by removing the shoot.

By contrast, changes that affect the lateral movement of auxin from cell to cell have major effects. If the lateral inward movement of auxin is reversed, the QC maximum vanishes as auxin is directed to the upward flows

of the superficial layers. If, instead, the inward reflux of auxin is increased, the slope of the QC-directed auxin gradient increases, and the total auxin content of the root increases. Instead of causing indigestion, reflux enables the root to behave like a capacitor, able to take up and maintain a graded charge of auxin without exogenous inputs.

Does this computerized auxin avatar have a real-life counterpart and, if so, what is it like? Grieneinsen et al.[2] use fluorescence-imaging methods on roots under a variety of conditions to show that auxin distributions correlate well with those predicted by cellular distributions of PIN proteins. Recognizing the morphogenic potential of a steep auxin gradient, the authors go on to link cellular behaviours observed in the root — including division, differentiation and expansion — to auxin concentration thresholds.

It is unclear exactly how an auxin gradient determines a range of distinct cellular behaviours. But the companion paper by Galinha et al.[3] highlights a role for transcription factors of the PLETHORA (PLT) group of proteins. In the model plant *Arabidopsis*, four genes that encode these proteins act to establish and maintain root pattern, and are expressed at both the messenger RNA and protein level in a gradient that closely parallels that of auxin. Based on changes in cellular behaviours that follow genetic perturbation of PLT levels, the authors propose that distinct cellular behaviours are evoked by the different concentration thresholds of PLT proteins, with PLT levels somehow being coupled to the PIN-mediated auxin gradient.

Many questions remain. Although auxin and PIN activities seem to reinforce each other, how are the geometrically precise PIN distributions across the root initially established? At a cellular level, asymmetric PIN localization can be understood in terms of polarized patterns of vesicle trafficking[10], but how are these patterns regulated? It is unlikely that they simply reflect intracellular auxin gradients, because the polarity of PINs can be dramatically reversed during early embryo formation[11]. Also unclear is whether, during these early stages, auxin gradients might depend more on other classes of transporter, or on mechanisms of auxin synthesis and turnover[12]. With respect to PLT levels providing a graded readout of auxin levels, how are these two coupled? The transcriptional induction of PLTs depends on derepression mediated through a particular class of auxin receptor, the F-box type of receptor, but it is unclear how direct this link is or whether other classes of receptor are also involved. Finally, what are the targets of PLT regulation, and how would their regulation by graded PLT levels determine distinct cellular behaviours?

Taken together, these two papers[2,3] provide a fresh look at patterning, reinforcing the potential role for auxin as a morphogen and providing a robust transport-dependent model for its graded distribution. They also illustrate the value of combining biologically grounded modelling with molecular analyses — an approach that is proving increasingly powerful for explaining biological complexity. ■
Bruce Veit is at AgResearch Grasslands, Tennent Drive, Private Bag 11008, Palmerston North, New Zealand.
e-mail: bruce.veit@agresearch.co.nz

1. Wolpert, L. *J. Theor. Biol.* **25,** 1–47 (1969).
2. Grieneisen, V. A., Xu, J., Marée, A. F. M., Hogeweg, P. & Scheres, B. *Nature* **449,** 1008–1013 (2007).
3. Galinha, C. *et al. Nature* **449,** 1053–1057 (2007).
4. Jones, A. M. *Science* **282,** 2201–2203 (1998).
5. Skoog, F. & Miller, C. O. *Symp. Soc. Exp. Biol.* **54,** 118–130 (1957).
6. De Smet, I. & Jürgens, G. *Curr. Opin. Genet. Dev.* **17,** 337–343 (2007).
7. Sabatini, S. *et al. Cell* **99,** 463–472 (1999).
8. Blakeslee, J. J., Peer, W. A. & Murphy, A. S. *Curr. Opin. Plant Biol.* **8,** 494–500 (2005).
9. Blilou, I. *et al. Nature* **433,** 39–44 (2005).
10. Geldner, N. *et al. Cell* **112,** 219–230 (2003).
11. Friml, J. *et al. Science* **306,** 862–865 (2004).
12. Quint, M. & Gray, W. M. *Curr. Opin. Plant Biol.* **9,** 448–453 (2006).

NUCLEAR PHYSICS

Neutrons cross the line

Paul-Henri Heenen

For most atomic nuclei, the maximum number of neutrons that can be bound is unknown. The discovery of two neutron-rich nuclei — and the confirmed absence of others — might help solve this conundrum.

Stars are element factories. Every atom in the Universe is a product of stellar nuclear reactions, but many of the stable isotopes that predominate today have formed from nuclei that existed only fleetingly. To understand the isotopic make-up of the Universe, we must first discover which of these ephemeral nuclei can form. In other words, what are the limits of nuclear composition? Reporting in this issue, Baumann et al.[1] describe their discovery of two new nuclei, one of which was thought to be impossible to make (see page 1022). This paves the way for a reassessment of nuclear models.

Atomic nuclei are composed of protons and neutrons, bound together by the strong nuclear force. Because this strong attraction is greater between a neutron and a proton than between two like particles, the most stable isotopes for light nuclei have equal numbers of neutrons and protons. But in larger nuclei — those with more than 20 protons — repulsion between the positively charged protons reduces the binding energy of the nucleus, so that the most stable isotopes have more neutrons than protons. In other words, the attraction between the extra neutrons counterbalances the electric repulsion between the protons.

There are about 300 stable nuclei, but these are only a small fraction of those that have a lifetime long enough to have a role in the formation of elements in stars. A chart of nuclei can be drawn that plots the number of protons against the number of neutrons (Fig. 1). Two lines — known as drip lines — can be plotted on the chart to indicate the limits of possible nuclei. On the proton-rich side of the chart, the proton drip line corresponds to the smallest number of neutrons that can be bound for a given number of protons. If a neutron were removed from a nucleus on the proton drip line, then a proton would be spontaneously emitted. Similarly, the neutron drip line indicates the largest number of neutrons that can accompany a given number of protons; extra neutrons will not bind to nuclei on this line.

The locations of the drip lines on the chart of nuclei are still an open question. Experiments performed in particle accelerators have proved the existence of about 3,000 isotopes[2]. Up to three times as many are thought to be awaiting discovery[3], but the short lifetimes of these bashful nuclei and the difficulty involved in making them makes this a formidable task. Nevertheless, some progress has been made. The location of the proton drip line is known for atoms with up to 90 protons — that is, up to thorium. This has been easier to establish than has the location of the neutron drip line, because electric repulsion between protons restricts the number that can be added to a nucleus with a given number of neutrons.

The situation is very different for neutron-rich nuclei, because no electric repulsion exists to limit the number of neutrons that can be added. The neutron drip line is therefore distant from the stable elements on the chart of nuclei, and so is much harder to reach experimentally. Furthermore, the number of neutrons that can theoretically be added to a nucleus increases as the number of protons increases.

The neutron drip line is therefore firmly established only up to oxygen, which has eight protons and can have a maximum of 16 neutrons. Beyond oxygen, the drip line has been tentatively assigned for elements up to sodium, which has 11 protons and a possible maximum of 26 neutrons. Baumann et al.[1] now report the discovery of two more neutron-rich isotopes: a magnesium isotope (^{40}Mg, which has 12 protons and 28 neutrons) and an aluminium isotope (^{42}Al, which has 13 protons and 29 neutrons).

The authors prepared these nuclei by firing the most neutron-rich calcium nuclei (^{48}Ca)
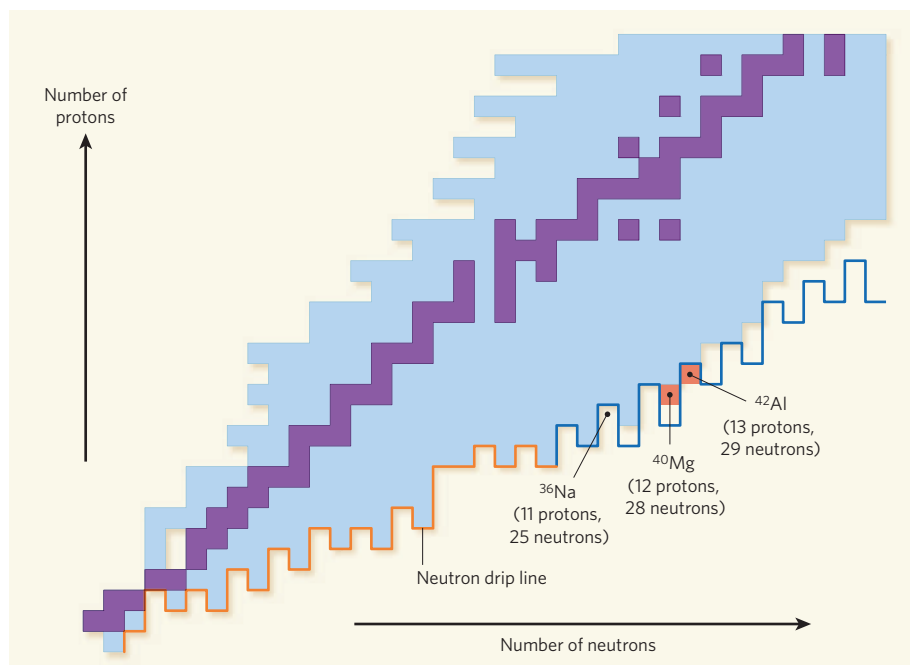
Figure 1 | Partial view of the nuclear landscape. This chart plots the number of protons against the number of neutrons for isotopes of the lightest elements. Purple squares represent stable isotopes, and pale blue squares represent those that have been observed only in particle accelerators. The orange line shows the experimentally determined 'neutron drip line', which indicates the largest number of neutrons that can be bound for a given number of protons. The blue line indicates the best predictions[6,7] of the neutron drip line for heavier elements. Baumann *et al.*[1] have now detected ^{40}Mg and ^{42}Al, but could not find evidence of ^{36}Na. These results suggest that the neutron drip line may extend further into the neutron-rich area than was expected.

at a natural tungsten target. Neutrons transfer from the tungsten to the ^{48}Ca nuclei, which can then break up to form ^{40}Mg or ^{42}Al. Although the number of nuclear interactions that could lead to the desired isotopes was extremely small (approximately 1 in 10^{15} reactions), Baumann *et al.* were able to identify them efficiently using state-of-the-art particle accelerator and detection technology. Just three collision events corresponded to the formation of ^{40}Mg, whereas 23 others led to that of ^{42}Al. The authors' results also seem to confirm that several other nuclei (^{30}F, ^{33}Ne and ^{36}Na) do not exist, as they were not observed during the experiment. This is curious, because isotopes of these three elements have been detected[4,5] that each contain a further neutron (that is, ^{31}F, ^{34}Ne and ^{37}Na).

The discovery of ^{42}Al and the non-appearance of ^{36}Na are telling results. These isotopes have odd numbers of both neutrons and protons. This means that they lack a stabilizing effect seen in other nuclei: superconductivity. In superconductors, electrons form pairs that differ only by the orientation of their spins. Protons and neutrons share the same tendency to pair up, with the nucleus receiving a boost of stability for each proton or neutron pair formed. An 'odd–odd' situation is a worst-case scenario for nuclei, because they have an unpaired proton and an unpaired neutron that do not contribute to superconductivity.

The idea of proton and neutron pairs helps explain irregularities in the neutron drip line. For example, ^{35}Na nuclei exist, but Baumann and colleagues' results suggest that it isn't

possible to add one neutron to make ^{36}Na. But it is possible to add two neutrons to produce ^{37}Na, an isotope with an even number of neutrons — presumably because the additional neutrons form a nucleus-stabilizing pair. Similarly, one can conclude from the stability of the odd–odd nucleus ^{42}Al that it should be possible to add an extra neutron to reach ^{43}Al, because the formation of a neutron pair will stabilize the heavier

isotope. In fact, the authors did observe a single collision event that is consistent with the existence of ^{43}Al, but this result requires further confirmation.

So why does ^{42}Al form at all, if it has an odd–odd nucleus? This isotope breaks the established staggered pattern of the neutron drip line, and was predicted not to exist. The answer could be that aluminium nuclei have reached a threshold of stability that minimizes the negative effect of extra neutrons, as has been suggested by some quantum-mechanical calculations. With the existence of ^{42}Al confirmed, there are now good arguments to suggest that the neutron drip line could extend as far as ^{47}Al.

Baumann and colleagues' results stretch the limits of current particle accelerators, so exploration of the neutron drip line at higher proton numbers will have to wait for the next generation of facilities. In the meantime, we must rely on models to predict where the neutron drip line falls — models that currently provide contradictory results. The authors' findings will certainly help to improve the accuracy of these models, so allowing a more confident assessment of the nuclear terra incognita. ■

Paul-Henri Heenen is at the Université Libre de Bruxelles, Physique Nucléaire Théorique et Physique Mathématique, CP 229, 50 Avenue F. D. Roosevelt, Brussels B1050, Belgium.
e-mail: phheenen@ulb.ac.be

1. Baumann, T. *et al.* Nature **449,** 1022–1024 (2007).
2. Audi, G., Wapstra, A. H. & Thibault, C. *Nucl. Phys. A* **729,** 337–676 (2003).
3. Lunney, D., Pearson, J. M. & Thibault, C. *Rev. Mod. Phys.* **75,** 1021–1082 (2003).
4. Notani, M. *et al. Phys. Lett. B* **542,** 49–54 (2002).
5. Lukyanov, S. M. *et al. J. Phys. G* **28,** L41–L45 (2002).
6. Möller, P., Nix, J. R., Myers, W. D. & Swiatecki, W. J. *Atomic Data Nucl. Data Tables* **59,** 185–381 (1995).
7. Samyn, M., Goriely, S., Bender, M. & Pearson, J. M. *Phys. Rev. C* **70,** 044309 (2004).

CANCER

Mixing cocktails

Charles L. Sawyers

A major hurdle in treating cancer is that tumour cells acquire drug resistance. To overcome this problem, one strategy might be to fine-tune the right mixture of drugs that target specific molecules.

Certain cancers are caused by oncogenic primary or 'driver' mutations in specific kinases — enzymes that regulate the activity of other proteins. Consequently, kinase inhibitors have been used in the clinic as effective single-agent drugs to shrink tumours. Kinase 'addiction' persists in advanced cancer, and patients who relapse after initially responding to kinase-inhibitor therapy often develop secondary mutations in the target kinase that confer drug resistance without impairing the kinase's oncogenicity[1]. Two reports[2,3] in *Science*

now show that lung cancers and glioblastoma — a malignant tumour of the central nervous system — use another option to escape drug treatment. Rather than mutate the drug target further, these cancers recruit other kinases that are not affected by the inhibitor to substitute for the pharmacologically impaired kinase and to restore downstream molecular signalling cascades that contribute to tumour growth.

Engelman and colleagues[2] stumbled across this new resistance mechanism by isolating cells from a lung-cancer cell line in which the driver
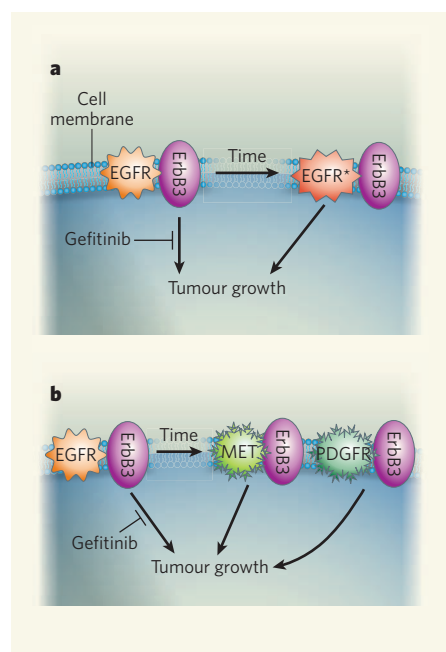
Figure 1 | The battle of tumours and drugs. Receptor tyrosine kinases (RTKs) have essential functions in mediating communication between a cell and its environment, and consequently activate downstream signalling cascades. One such RTK is EGFR, which following a primary, or driver, mutation couples with another RTK, ErbB3, promoting tumour-cell growth. EGFR-inhibitor drugs, such as gefitinib, block the oncogenic activity of EGFR, leading to tumour shrinkage. However, with time, tumour cells become resistant to RTK inhibitors by one of two mechanisms. **a,** It is well established that drug resistance can be caused by a secondary mutation in EGFR (EGFR*), which prevents drug binding and restores tumour growth. One way to tackle this might be to develop second-generation EGFR inhibitors. **b,** Two new studies[2,3] reveal an alternative mechanism of drug resistance. They show that, although gefitinib inhibits EGFR, other RTKs, such as MET or PDGFR, which are unaffected by this drug, might instead couple with ErbB3 and take over the task of tumour regrowth. In this case, a cocktail of drugs to simultaneously inhibit EGFR, MET and PDGFR is required.

mutation is in a kinase known as EGFR. This belongs to a group of kinases known as receptor tyrosine kinases (RTKs). The authors found that these cells, which initially responded to an EGFR-inhibitor drug, gefitinib, subsequently became resistant, but the resistance was not due to a secondary mutation in EGFR. Instead, in these cells, the levels of another RTK, MET, were unusually high owing to amplification of the gene that encodes it. The researchers then inhibited MET activity and found that, in the presence of EGFR inhibitors, this treatment restored the cancerous cells' drug sensitivity — an observation that makes a case for combination therapy in the treatment of drug-resistant forms of lung cancer.

Before this work, the only known cause of resistance to EGFR inhibitors in lung cancer

was a secondary mutation — from threonine to methionine — at position 790 in the amino-acid sequence of EGFR, abbreviated to T790M (ref. 4). The T790 residue is evolutionarily conserved among kinases and is implicated in drug resistance with other kinase–inhibitor pairings in other cancers. The T790M mutation occurs in more than half of patients with non-small-cell lung cancer who relapse after treatment with gefitinib or another EGFR-inhibitor drug, erlotinib. Engelman *et al.*[2] suggest that amplification of MET activity may account for 20% of the remaining cases of resistance.

The work of Engelman *et al.* highlights the principle that, in cancer, signal-transduction pathways are remarkably flexible and amenable to rewiring. In this case, MET doesn't overcome EGFR inhibition by simply driving the canonical MET-mediated signalling pathways. Instead, it undergoes a form of molecular morphing, whereby it becomes an illicit partner of the EGFR co-receptor ErbB3 and acquires the signalling properties of EGFR (Fig. 1).

If resistance options are viewed from the perspective of a cancer cell, this RTK switch mechanism seems an easy strategy by which to escape drug inhibition, compared with the evolutionary work involved in selecting a rare second-site mutation that would impair drug activity without crippling the oncogenicity of the original RTK. Yet, so far, second-site mutations are much more commonly documented. This presumably reflects the fact that very few tumour cell types have the molecular circuitry to allow RTK switching. Mathematically, the probability of generating secondary mutations is low[5], but the probability of acquiring the still-unknown genetic makeup required to rewire a signal-transduction pathway must be lower.

In another study, Stommel and colleagues[3] describe a conceptually related RTK switch in glioblastoma. As in the lung study[2], the authors show that MET can substitute for EGFR function, but with a new twist. They find that the tumour cells are resistant to EGFR inhibitors upfront (rather than acquiring resistance after prolonged drug exposure), but become sensitive when MET and EGFR are inhibited in parallel. So it is as if these tumours have already been through the cycle of drug response and relapse described by Engelman *et al.*, when in reality they have never been exposed to these drugs. Furthermore, inhibition of a third RTK — PDGFR — is required for maximal drug efficacy. It therefore seems as though these cancers have conspired to implement several back-up systems against single-agent kinase-inhibitor treatment.

Cancer cells are admittedly wily, but the more likely interpretation of this behaviour is that the orchestra of RTKs collectively contributes to disease progression. Each player presumably makes an individual contribution, but can pick up the slack if one partner is crippled. Indeed, further analysis of acquired resistance in lung

50 YEARS AGO

In his presidential address to the Library Association … Dr. J. Bronowski said that a civilized society must preserve what its best minds discover, but preservation alone does not make it an educated or even a cultured society. Moreover, an educated society could exist only when knowledge is not merely stored but is also shared, and it was the invention of printing that made the book an instrument of education … but in science the public libraries have scarcely played that part at all: if they are to do so, they must have the books to enable them to make the language of science familiar to those who are not professional scientists … He did not think that the printed book was the last instrument of education we would discover, but he was sure that the printed book and the public library would remain the most powerful means of self-education.
From Nature 26 October 1957.

100 YEARS AGO

In India at the present moment the ravages of plague, though not so great as those of the Black Death or of the Great Plague in London, are nevertheless dreadful. During the first six months of this year no less than 1,060,000 deaths from plague occurred in India, and out of these 632,000 occurred in the Punjaub, which has a population of only twenty-five millions, that is to say, one in every forty inhabitants in this district has died of plague between January and June … The great difficulties in the way of preventative measures are ignorance and apathy, to which superstition is often superadded. In some parts of India there is great prejudice against taking life of any kind … Cases of plague from time to time arrive at the port of London … We are pursuing a foolish policy in allowing rat- and flea-infected districts to exist in the East End of London and other similar places.
From Nature 24 October 1907.

50 & 100 YEARS AGO

cancer indicates a similar scenario, whereby MET amplification and mutation in the T790 residue of EGFR coexist in patients and in cell lines that have never been exposed to EGFR-inhibitor drugs (J. Bean and W. Pao, personal communication). Together with another study[6] showing that some drug-resistant forms of the cancer-associated kinase BCR–ABL have enhanced oncogenicity, there is growing evidence that these drug-resistant proteins also have a role in tumour progression.

The most important implication of the work of Stommel et al.[3], which is yet to be clinically tested, is that a much larger spectrum of cancers, not just those with driver kinase mutations, might be sensitive to cocktails of kinase inhibitors if such mixtures are delivered in the right combination. But how will we recognize these multiple-kinase-dependent cancers? Borrowing from the approach used by Engelman et al.[2] and Stommel and colleagues[3] to uncover the RTK-switch phenomenon, one possibility is to look at the cellular profile of RTK activation in tumour biopsies, thereby distinguishing between single-kinase-dependent and multiple-kinase-dependent cancers.

Technologies for conducting extensive profiling of proteins that carry signs of kinase activity in tumour cells, such as mass spectrometry and antibody-based approaches, are readily accessible. So 'signatures' of RTK activation in tumour biopsies could complement information gained through analysis of gene mutation status — information that is currently used to define tumours with mutations in genes encoding kinases.

Preclinical strategies to estimate the fraction of multiple-kinase-dependent cancers might also be feasible using large panels of tumour-cell lines. Such high-throughput, cell-based screens are already gaining popularity as a tool to determine the relationship between genotype and response to single-drug inhibitors. But, within reason, they could be expanded to evaluate the genotype relationship with multiple-kinase inhibition.

The new findings[2,3] add fuel to the argument that, using rationally chosen combinations of kinase inhibitors, successful cancer treatment can be achieved. Prior to these reports, this view had originated from the knowledge that second-site kinase mutants are a principal cause of single-agent resistance in tumours with driver mutations in kinases. Therefore, second-generation compounds must be effective against these drug-resistant proteins and should be used in combinations that prevent the emergence of resistant subclones. A good example is the proposed combined use of the ABL-kinase-inhibitor drugs imatinib and dasatinib to treat chronic myeloid leukaemia[7–9]. The kinase-switching phenomenon reported by Engelman et al. and Stommel et al. requires a broader view that incorporates drugs that target relevant bystander RTKs. So, clearly,

the cocktail menu will continue to grow, but the cocktails should be mixed with appropriate molecular guidance. ∎

Charles L. Sawyers is at the Memorial Sloan-Kettering Cancer Center, Human Oncology and Pathogenesis Program, 1275 York Avenue, New York, New York 10065, USA.
e-mail: sawyersc@mskcc.org

1. Sawyers, C. L. Genes Dev. 17, 2998–3010 (2003).
2. Engelman, J. A. et al. Science 316, 1039–1043 (2007).
3. Stommel, J. M. et al. Science 318, 287–290 (2007).
4. Pao, W. et al. PLoS Med. 2, e73 (2005).
5. Michor, F. et al. Nature 435, 1267–1270 (2005).
6. Skaggs, B. J. et al. Proc. Natl Acad. Sci. USA 103, 19466–19471 (2006).
7. Shah, N. P. et al. Cancer Cell 2, 117–125 (2002).
8. Shah, N. P. et al. J. Clin. Invest. 117, 2562–2569 (2007).
9. Talpaz, M. et al. N. Engl. J. Med. 354, 2531–2541 (2006).

POPULATION ECOLOGY

Group living and hungry lions

Tim Coulson

Ecologists have necessarily had to simplify matters in looking at predator–prey dynamics. Study of a situation in which predator and prey live in groups reveals that a key process was previously overlooked.

Eating is a necessity of animal life. So you might expect that individuals would do everything possible to maximize their food intake. An example, often invoked, is that lions live in prides because group hunting increases food availability by allowing the lions to kill prey that would be too large for an individual to tackle. On page 1041 of this issue, however, Fryxell et al.[1] argue that living in groups decreases the amount of food each lion eats. Moreover, that decrease is even more severe if the lions' prey is also gregarious. How can this be explained?

Many large predators live in groups, as is dramatically seen in the familiar TV programmes that show, for example, a pack of wild dogs running down an impala or a pod of killer whales attacking a shoal of herring. Many of the prey species of these top carnivores also

live in groups (Fig. 1). But there has been surprisingly little research into how group living influences individual food-intake rates, and the dynamics of the populations of predators and their prey.

A concept called the functional response[2] lies at the heart of behavioural and population ecology, and of Fryxell and colleagues' paper. This is the curve that describes the intake rate of a single predator as a function of prey density. The shape of the curve is the result of two processes — the rate at which predators encounter prey and the speed with which they consume it[3]. Much of the theory investigating the consequences of different shapes of functional response has been developed following observations and experiments from systems where both predator and prey are solitary. These systems often show dynamics in which

Figure 1 | Group theory. Both the wildebeest and the lioness pictured here are gregarious animals, but Fryxell and colleagues[1] conclude that the group life of both species doubly diminishes the food intake of an individual lion.

the numbers of both predator and prey increase and decrease, sometimes leading to extinction of first prey and then predator. Altering the form of the functional response can, in some circumstances, prevent extinction[4].

Fryxell *et al.*[1] examine how group living in prey, in predators and in both kinds of species influences the shape of the functional response and the interaction between predator and prey populations. They show theoretically that gregarious living in either the prey or the predator species reduces the rate of prey consumption by each predator. Intake rates are lowest when both species live in groups. If prey lives singularly, an increase in the number of prey will lead to a linear increase in the likelihood that a predator will encounter a meal on its daily wanderings. However, if the prey forms clumps, there will be large holes in the landscape through which a predator can roam without finding something to eat. Group living in prey therefore decreases intake rate. When a predator does find a group of prey, it encounters an embarrassment of riches and quickly becomes satiated. The intake rate of predators is reduced if they live gregariously because each individual searches the same area and then has to share the prey that it kills.

Fryxell *et al.* primarily use data from surveys of lions preying on wildebeest in the Serengeti National Park, Tanzania, to reach their conclusions for a system where both predator and prey live gregariously. They show that the consequence of both species living in groups is a predicted reduction in the food-intake rate per lion of 90% compared with the rate when lions forage solitarily. This is an enormous amount, and is equivalent to the decrease in food availability that results if the migratory wildebeest is present in a lion-pride territory for only a fraction of each year. The authors also examine the consequence of such a large reduction in intake rate on the dynamics of the lion and wildebeest populations. They predict that, if both species shunned group living and wandered the plains singly, the dynamics of both populations would be highly unstable, with both predator and prey likely to become extinct. In contrast, if both predator and prey live in groups, it is much more likely that both their populations will persist.

But the question of why lions live in groups remains. Fryxell *et al.* argue that the benefits primarily accrue from territory defence and the communal protection of young against males that can kill cubs when they take over a pride[5,6]. However, an argument familiar to most ecologists is that lions live gregariously because group hunting is required to bring down large prey. Fryxell *et al.* accept that group hunting does allow lions to attack and kill the formidable Cape buffalo, but they also state that: "Most individual lions refrain from contributing to group hunts." This carefully worded statement does not rule out the possibility that substantial benefits arise from group hunting, and it flags one of the problems of parametrizing functional responses. It is challenging to work

out individual intake rates accurately across different sizes of predator and prey groups for a range of prey densities, especially for species with such complex social arrangements as those seen in lions. I would be interested to know whether functional responses derived entirely from observations on the feeding of individual lions would yield similar conclusions to those obtained by Fryxell *et al.* using survey data.

This work[1] shows that extending the functional response to include some realistic natural history helps to explain why the inevitable extinction of predators and prey, predicted by some simple population models, is not observed in the wild. The paper will stimulate researchers who have obtained functional responses using detailed observational data on group-living predators. And it should encourage theoreticians to examine how other aspects of animal behaviour might affect the predictions derived from simple population models. ■

Tim Coulson is in the Department of Life Science, Silwood Park Campus, Imperial College London, Ascot, Berkshire SL5 7PY, UK.
e-mail: t.coulson@imperial.ac.uk

1. Fryxell, J. M., Mosser, A., Sinclair, A. R. E. & Packer, C. *Nature* **449,** 1041–1043 (2007).
2. Holling, C. S. *Can. Entomol.* **91,** 293–320 (1959).
3. Jeschke, J. M., Kopp, M. & Tollrian, R. *Ecol. Monogr.* **72,** 95–112 (2002).
4. Abrams, P. A. & Ginzburg, L. R. *Trends Ecol. Evol.* **15,** 337–341 (2000).
5. Schaller, G. B. *The Serengeti Lion: A Study of Predator-Prey Relations* (Univ. Chicago Press, 1976).
6. Packer, C., Scheel, D. & Pusey, A. E. *Am. Nat.* **136,** 1–19 (1990).

ATTOPHYSICS

At a glance

David M. Villeneuve

Measurements on the attosecond timescale had been limited to the dynamics of electrons in an atomic gas. But a record has now been set in a quite different context — the photoemission of electrons from a surface.

The quest for faster and faster time-resolved measurements has reached a new level: Cavalieri *et al.* report (page 1029 of this issue)[1] that they have measured a delay of 100 attoseconds in the emission of electrons ejected from a surface irradiated by light. This is not just the experiment with the best time resolution yet; it is also the first time that attosecond metrology has been applied to a solid, rather than a gaseous, system.

It was only in the 1990s that the trend to ever faster measurements produced laser sources with pulse durations below 5 femtoseconds (a femtosecond is 10^{-15} seconds). This is the timescale of the motion of atoms within molecules. Femtochemistry, in which a chemical reaction is followed through its transition state, became big news[2]. But that is now old hat. The attosecond (10^{-18} seconds) is the timescale of the motion of electrons within atoms: an electron takes about 150 attoseconds to orbit a hydrogen atom.

Attosecond pulses are created when intense laser pulses of femtosecond duration are focused into a gas sample. A process known as high-harmonic generation[3] then kicks in to produce light at a range of frequencies that are precisely phased together, creating a train of very short, coherent pulses. In the past few years, the technology has evolved to the point where single pulses just 130 attoseconds long can be produced with table-top-sized laser systems[4]. These pulses are so short that their frequency (and thus energy) lies in the extreme-ultraviolet or soft-X-ray portion of the electromagnetic spectrum.

Attosecond metrology has previously been applied to samples of atomic gases to observe excitation processes of electrons such as shake-up and Auger decay[5]. These are essentially 'pump–probe' measurements: an attosecond pulse excites the system, and the intense optical laser field that generated the pulse follows it and is used to sweep up the charged products — much as an oscilloscope streaks an electron beam across the screen to resolve an electrical pulse.

Cavalieri *et al.*[1] focus their 90-electronvolt extreme-ultraviolet pulse at an angle on a tungsten metal surface (Fig. 1, overleaf). The lower-frequency optical pulse that created the attosecond pulse follows along the same path, but its passage can be delayed in steps of 300 attoseconds. Electrons liberated through the photoelectric effect by the first pulse are detected by a spectrometer that measures their kinetic energy. The optical laser field pushes these photoelectrons' energy up or down, depending on the precise position of both the attosecond pulse and the photoelectron in the laser field's cycle.

By varying the time delay between the pulse and the optical field, and measuring the shift in the up and down motion of the energy spectrum, the authors could precisely measure the emission time of the photoelectrons. They were able to distinguish electrons coming from different energy states in the surface, observing that electrons from the more deeply bound core states in the surface were emitted
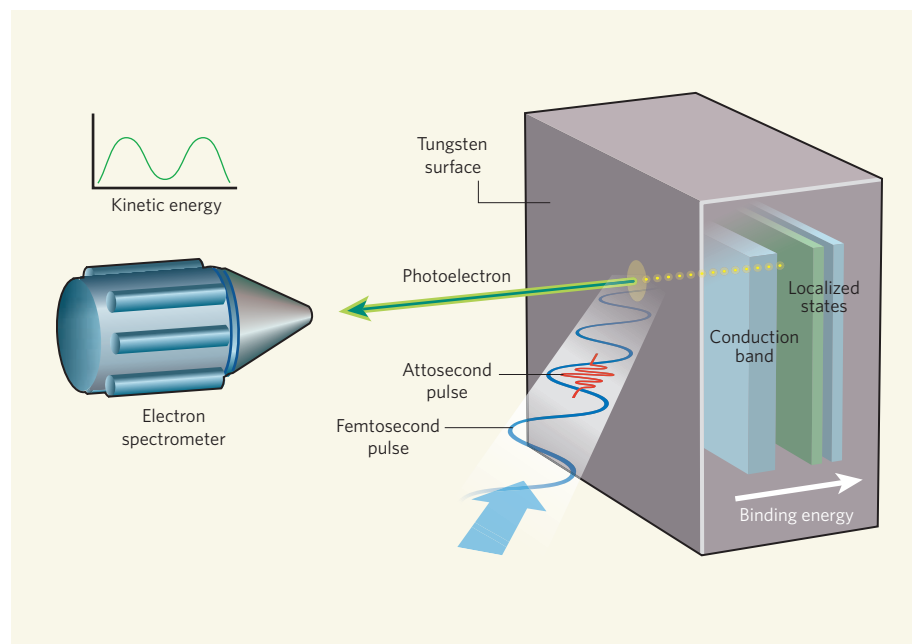
Figure 1 | Fast light on a dark place. In Cavalieri and colleagues' set-up[1], an extreme-ultraviolet, attosecond pulse is produced from an optical femtosecond laser field, and dislodges electrons from a tungsten surface through the photoelectric effect. The kinetic energy of the electrons, measured in a spectrometer, determines the energy level from which they were ejected. The authors observe a delay of around 100 attoseconds between electrons emitted from the energetically shallow conduction band of the metal and those emerging from deeper localized states.

around 100 attoseconds after those from the conduction band.

The authors' great triumph is to apply the streaking technique to a metal surface, rather than a gas sample. This required an optical field of intermediate strength — intense enough to modify the energy of the photoelectrons significantly, but not so intense as to create photoelectrons directly.

When it comes to sources of soft X-rays of high brightness, synchrotrons still reign supreme. But the pulse duration of synchrotron light is in the picosecond range, and thus much too slow to observe the dynamics of fast electrons. Next-generation free-electron-laser facilities, such as FLASH at the DESY research centre in Hamburg, Germany, have pulse durations of around 50 femtoseconds, and will allow images to be obtained of molecular structures during chemical reactions, or protein structures to be determined[6]. The final kilometre of the SLAC linear accelerator at Stanford University is also being converted into a source of femtosecond X-ray pulses, the Linac Coherent Light Source, or LCLS.

But quite apart from the reduced cost, attosecond pulses produced from tabletop lasers have one significant advantage over these very expensive pulsed-X-ray facilities — they are exquisitely synchronized with the laser pulse that produces them. In our facility in Ottawa, we have found that attosecond pulses produced in two separate sources by the same laser are synchronized to even better than an attosecond — in the zeptosecond (10^{-21} second) range. By contrast, free-electron sources have significant timing 'jitter' that limits their resolution in pump–probe experiments. And, as any physicist living out of a suitcase will tell you, working in one's own laboratory is far preferable to travelling to a distant facility. ∎

David M. Villeneuve is at the National Research Council of Canada, 100 Sussex Drive, Ottawa, Ontario K1A 0R6, Canada.
e-mail: david.villeneuve@nrc.ca

1. Cavalieri, A. L. *et al. Nature* **449,** 1029–1032 (2007).
2. Zewail, A. H. http://nobelprize.org/nobel_prizes/chemistry/laureates/1999/zewail-lecture.html (1999).
3. Scrinzi, A., Ivanov, M. Yu., Kienberger, R. & Villeneuve, D. M. *J. Phys. B* **39,** R1–R37 (2006).
4. Sansone, G. *et al. Science* **314,** 443–446 (2006).
5. Uiberacker, M. *et al. Nature* **446,** 627–632 (2007).
6. Neutze, R. *et al. Nature* **406,** 752–757 (2000).

BIOCHEMISTRY

Indifferent chaperones

Eckhard Jankowsky

How do nonspecific enzymes that help to correct RNA folding identify misfolded structures among similar, properly folded RNAs? It seems that careful discrimination has little to do with it.

Most RNAs adopt defined three-dimensional structures that allow them to function properly. But RNAs also often misfold into inactive structures, which can persist for a long time[1]. Fortunately, RNA chaperone proteins give misfolded RNAs the opportunity to refold correctly[1,2]. However, many RNA chaperones are nonspecific, and correctly folded and misfolded RNA structures tend to be very similar. On page 1014 of this issue, Bhaskaran and Russell[3] describe the way one RNA chaperone, CYT-19, confronts this challenge.

CYT-19 is a member of the large group of DEAD-box proteins — enzymes that are involved in almost all aspects of RNA metabolism[4]. The name DEAD-box comes from the amino-acid sequence (D-E-A-D in single-letter code) of one of the evolutionarily conserved motifs in these proteins. DEAD-box proteins are often called RNA helicases, because they can unwind short RNA duplexes using ATP as an energy source. But in contrast to canonical helicase enzymes, which unwind duplexes by travelling along the nucleic-acid strands, DEAD-box proteins do not move on the RNA duplexes but load directly to double-stranded regions and then pry the strands apart[5].

Among the targets of CYT-19 are misfolded group-I intron RNAs, which are catalytic RNAs that excise themselves from, and subsequently rejoin the ends of, a larger precursor RNA. It is thought that CYT-19 unwinds one or more of the short helices that are the building-blocks of larger RNA structures[6]. Like most DEAD-box helicases, CYT-19 unwinds duplexes indiscriminately and, given that the correctly folded and misfolded group-I introns form almost identical helices, it was not clear how CYT-19 distinguishes between these structures. The general assumption was that, somehow, CYT-19 unravels only the misfolded structures to promote correct folding.

Bhaskaran and Russell[3] make the remarkable observation that CYT-19 unfolds both correctly folded and misfolded structures. However, misfolded RNAs are disassembled at a faster rate. Consequently, at any given time, the misfolded group-I intron RNAs have a better chance of unfolding and refolding correctly (Fig. 1, overleaf).

But why are misfolded RNAs disassembled faster than the correct structures? The authors show that mutated group-I introns that are functional but cannot pack all their helices together in the ultimate higher-order

structure are readily unfolded by CYT-19. So it is possible that the reason for the slower unfolding rate of correctly folded RNAs is that crucial helices are sequestered in the higher-order structure and are thus protected from unwinding by CYT-19. This possibility is supported by an earlier study[7] showing that CYT-19 cannot easily unwind duplexes hidden within an RNA structure. If this is true for other nonspecific DEAD-box chaperones, this would be an elegant and unexpectedly simple way by which these proteins might preferentially unravel misfolded RNA structures.

Bhaskaran and Russell also make another intriguing observation. Under certain conditions — for example, at reduced magnesium concentrations, a consequence of which is the weakening of higher-order RNA structures — CYT-19 increases the ratio of misfolded RNAs to correctly folded structures. This is remarkable, because misfolded group-I RNAs are thermodynamically less stable than the correct structures and, at equilibrium, the most stable species dominates in a mixture of possible structures.

The authors find that CYT-19 achieves the redistribution of RNA species against their thermodynamic equilibrium by establishing a 'kinetically controlled steady state' (Fig. 1b). In this state, CYT-19 constantly unwinds RNA structures in an ATP-dependent manner, so that the distribution of these structures is no longer dictated by their thermodynamic stabilities but, instead, by the rate at which they are actively unravelled by CYT-19, as well as by the rate of their refolding. Under these conditions, CYT-19 readily unravels both misfolded and correct structures. But when the RNA subsequently misfolds it does so faster than correctly refolding RNA, and therefore accumulates.

Similar ATP-driven redistribution of model RNA complexes by another DEAD-box protein has been reported[8]. So it seems that DEAD-box proteins can establish kinetically controlled steady states for various RNAs. The ability of these proteins to favour the accumulation of thermodynamically less stable RNA species seems particularly beneficial for processes in which RNAs must undergo extensive structural changes, such as during ribosome assembly and pre-messenger-RNA splicing. Both of these processes involve numerous DEAD-box proteins[4].

The work of Bhaskaran and Russell[3] implies that sequence nonspecificity of DEAD-box proteins is more of an asset than a disadvantage, at least for their function as RNA chaperones. Nonspecific chaperone proteins can work on many misfolded RNAs and at different positions within them, which obviates the need for a large number of highly specific proteins that would exclusively aid the folding of each of the many cellular RNAs. The potency of unfolding by the omnipresent and nonspecific DEAD-box proteins also offers one possible explana-
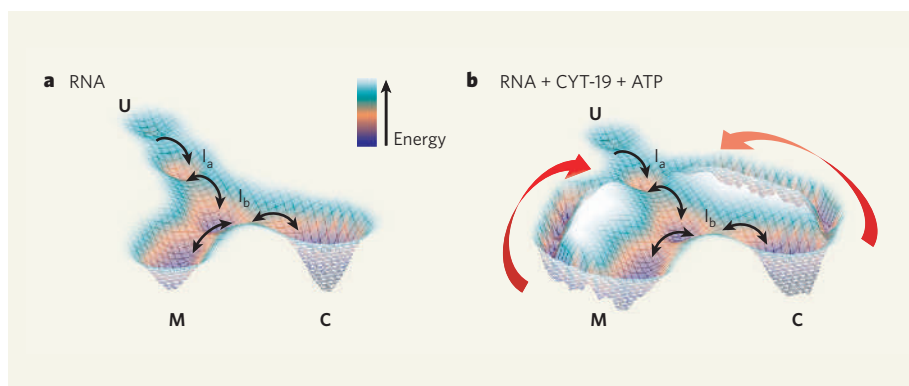

Figure 1 | Schematic folding landscape of group-I intron RNAs. In this representation, valleys indicate structures of varying stability — the deeper a valley, the lower the energy of the structure and the more stable it is. The ridges between valleys correspond to rates of inter-conversion between neighbouring structures; high ridges indicate slow rates. U represents an unfolded RNA, I_a and I_b are intermediate structures, M is a misfolded and C a correctly folded RNA. **a,** Left to their own devices, most group-I intron RNAs first form M, which is less stable than C but forms more quickly. However, M converts back into I_b and I_a at an appreciable rate, allowing further rounds of folding into both C and M. Although C can also convert back to I_b, this process is slow. So with time, C accumulates. **b,** Bhaskaran and Russell[3] show that CYT-19 alters the RNA-folding landscape in an ATP-dependent manner. By unfolding both M and C, new pathways for structural conversions are opened that presumably lead to I_a/I_b (red arrows). The formation of the RNA species depends on the rates at which they can interconvert. This represents a kinetically controlled steady state, in which less stable structures can accumulate. The levels of ATP determine and maintain the topology of the folding landscape, and in the absence of ATP the landscape changes back to that seen in **a.**

tion for why some RNAs tolerate misfolding into fairly stable structures. This is because they can rely on DEAD-box proteins to rescue them. But an increased awareness of the versatility of nonspecific DEAD-box proteins highlights the next question: why does the cell use so many of these enzymes? ∎

Eckhard Jankowsky is in the Department of Biochemistry and the Center for RNA Molecular Biology, School of Medicine, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA.
e-mail:exj13@case.edu

1. Herschlag, D. *J. Biol. Chem.* **270,** 20871–20874 (1995).
2. Schroeder, R., Barta, A. & Semrad, K. *Nature Rev. Mol. Cell Biol.* **5,** 908–919 (2004).
3. Bhaskaran, H. & Russell, R. *Nature* **449,** 1014–1018 (2007).
4. Linder, P. *Nucleic Acids Res.* **34,** 4168–4180 (2006).
5. Yang, Q. & Jankowsky, E. *Nature Struct. Mol. Biol.* **13,** 981–986 (2006).
6. Mohr, S., Stryker, J. M. & Lambowitz, A. M. *Cell* **109,** 769–779 (2002).
7. Tijerina, P., Bhaskaran, H. & Russell, R. *Proc. Natl Acad. Sci. USA* **103,** 16698–16703 (2006).
8. Yang, Q., Fairman, M. E. & Jankowsky, E. *J. Mol. Biol.* **368,** 1087–1100 (2007).

ENVIRONMENTAL SCIENCE

Nutrients in synergy

Eric A. Davidson and Robert W. Howarth

A literature meta-analysis of the effects of nitrogen and phosphorus on plant growth prompts a thought-provoking inference — that the supply of, and demand for, these nutrients are usually in close balance.

The nutritional needs of plants start with carbon, hydrogen and oxygen, which they obtain from atmospheric gases and from water. They also require nitrogen (a constituent of all proteins) and phosphorus (not least as a component of nucleotides, including those in DNA and RNA). Writing in *Ecology Letters*, Elser and colleagues[1] provide a timely contribution to our understanding of how nitrogen and phosphorus, both individually and in combination, affect primary producers such as crops, trees and algae.

This is a topic with a long history. In the nineteenth century, Justus von Liebig proposed his 'law of the minimum', which describes crop production as being limited by the nutrient in shortest supply. According to this law, once the nitrogen (N) need of a crop is met by fertilization, another element such as phosphorus (P) becomes limiting (Fig. 1a), and there is no further response to additional N.

Since Liebig's day, the use of N and P fertilizers has revolutionized agriculture. The consequences have been both desirable and

undesirable. Synthetic fertilizers have fuelled the Green Revolution and greatly lessened world hunger and malnutrition. But they have also led to severe air and water pollution and other adverse effects on the environment and human health[2,3]. The widespread distribution of excess N and P within natural ecosystems has caused degradation of coastal waters globally, and an increase in hypoxic areas ('dead zones') on nearly every continent[3,4]. Excess N and/or P lead to algal blooms and so consumption of the oxygen required for productive fisheries and for healthy marine habitats. In the United States, two-thirds of coastal rivers and bays are degraded through the consequences of excess nutrients[5]. Similar problems occur in fresh waters, and excess fertilization of forests can lead to decline and loss of biodiversity. Management strategies for mitigating the two pollutants can differ[2]. So the long-standing debate over the relative importance of N and P as agents of excess production in different ecosystems is of immense practical significance.

This is the context in which Elser and colleagues' study[1] is set. It is a meta-analysis of more than 300 publications reporting results of nutrient-amendment experiments in marine, freshwater and terrestrial ecosystems, with the effects of the two elements being assessed in terms of increased biomass or production. The study supports some well-established rules of thumb in biogeochemistry. Examples are the greater limitation on P than N in mature forests growing on highly weathered lowland soils in the tropics; the greater responses to P than N addition in freshwater ecosystems; and the greater responses to N addition in marine ecosystems.

More importantly, the analysis demonstrates a surprisingly consistent pattern of a synergistic effect of N and P addition on net primary productivity across all ecosystem types. Adding N and P together seems to give photosynthesis by algae and higher plants more of a boost than adding either one separately. The authors infer from this that the stoichiometry of N and P supply and demand must generally be in close balance in most ecosystems. According to this interpretation, P is rarely available in great excess relative to N, so a modest addition of N quickly provokes a limitation on P. When N and P are added together, N and P limitation may alternate in numerous small incremental steps, ultimately producing a synergistic effect (Fig. 1b).

This is an impressive synthesis[1], but several caveats are in order. First, we lack a mechanistic understanding of how the availability of one resource affects the supply of and demand for another resource. At the cellular level, regulation of the relative dynamics of the demand for N for the synthesis of enzymes, and the demand for P for the synthesis of nucleic acids and also of ATP, is poorly known. At the organismal level, some species have adaptations for obtaining a nutrient that would otherwise be difficult to acquire (for example, diverting carbon and
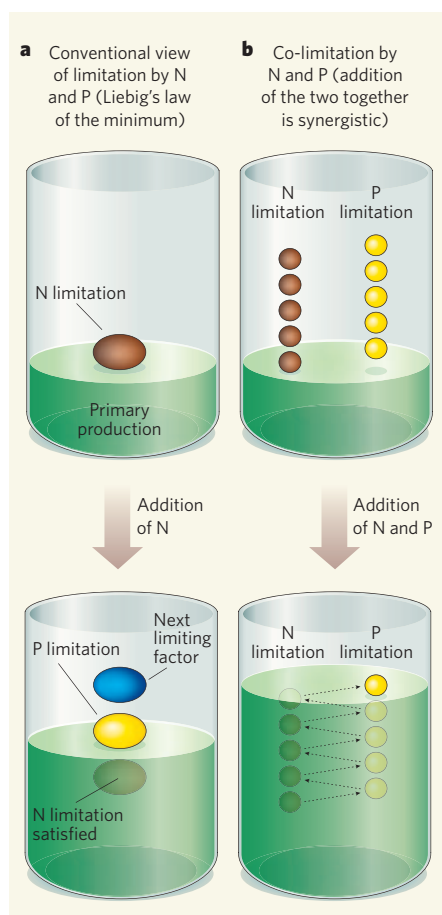

Figure 1 | Nutrients and primary production. a, The conventional view of Liebig's law of the minimum. Here, after the need for nitrogen (N) is met, N is not limiting, and increased production successively depends on satisfying limitation by phosphorus (P) and then any other limiting factors. **b,** A modification of the law of the minimum, based on one interpretation of the synthesis by Elser *et al.*[1]. From their meta-analysis, they infer that supply and demand of N and P are usually in close balance, so that an incremental addition of one causes a modest limitation by the other. Alleviation of each incremental limitation in turn produces a synergistic effect when both nutrients are added together.

nitrogen to the fungal mycorrhizae that form symbioses with plant roots to improve access to P in nutrient-poor soils). But our knowledge of such trade-offs is only qualitative. And at the ecosystem level, the factors that may influence the relative importance of N fixation from the atmosphere, such as the availability of N, P and molybdenum (an essential component of N-fixing enzymes), and competition for light and water, are not well understood[6].

Second, the doses of N and P addition — either through natural pulses or in bioassay experiments — may be very important, but were not specifically addressed by Elser and colleagues. If the doses are large enough, adding both nutrients may simply alleviate first one nutrient limitation and then the other,

as per Liebig's classic law, and this could look like a synergistic effect. The assertion that the stoichiometric supply of N and P in natural systems is generally close to balanced could be further tested by quantitative dose–response experiments.

Third, there is the issue of the timescale of nutrient-amendment experiments. Experiments that are short in duration relative to the life cycles of the organisms being studied measure only the response of the organisms that are dominant in the ecosystem at the time of the assay[5,7]. Longer-term ecosystem-scale responses to nutrients can be different, as the dominant species change with changes in biogeochemical processes. A short-term assay in a freshwater lake, for example, would probably indicate P limitation, followed by an apparent synergistic effect with N limitation once enough P were added. But in one experiment[8], years of P fertilization led to a predominance of N-fixing cyanobacteria that produced enough reactive nitrogen to keep the lake P limited.

Finally, resource limitation may simultaneously involve several nutritional elements, along with light, water and carbon dioxide. If responses to the addition of N and P are synergistic, we would expect complex synergies among other potential limiting resources as well.

So we are left with plenty of questions. Nonetheless, Elser and colleagues' meta-analysis[1] provides the most thorough examination of short- and medium-term nutrient-amendment experiments to date. Their synthesis adds to a growing body of evidence that addressing the 'off-farm' environmental consequences of food production will require efforts to reduce losses of both N and P from agricultural systems[2,5,7]. It also provides an improved basis for formulating testable hypotheses to describe nutrient interactions in other ecosystems. ■

Eric A. Davidson is at the Woods Hole Research Center, 149 Woods Hole Road, Falmouth, Massachusetts 02540-1644, USA. Robert W. Howarth is in the Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, New York 14853, USA.
e-mails: edavidson@whrc.org;
rwh2@cornell.edu

1. Elser, J. J. *et al. Ecol. Lett.* doi: 10.1111/j.1461-0248.2007.01113.x (2007).
2. Howarth, R. W. *et al.* in *Millennium Ecosystem Assessment. Ecosystems and Human Well-being* Vol. 3: *Policy Responses* 295–311 (Island, Washington DC, 2005).
3. UNEP/WHRC. *Reactive Nitrogen in the Environment: Too Much or Too Little of a Good Thing* (United Nations Environment Programme, Paris, 2007). www.whrc.org/policy/Reactive_nitrogen.htm
4. Diaz, R. J., Nestlerode, J. & Diaz, M. L. *Proc. 7th Int. Symp. Fish Physiology, Toxicology, and Water Quality* EPA-600-R-04-049, 1–33 (US Environmental Protection Agency, 2003).
5. National Research Council. *Clean Coastal Waters: Understanding and Reducing the Effects of Nutrient Pollution* (National Academies Press, Washington DC, 2000).
6. Vitousek, P. M. *et al. Biogeochemistry* **57/58,** 1–45 (2002).
7. Howarth, R. W. & Marino, R. *Limnol. Oceanogr.* **51,** 364–376 (2006).
8. Schindler, D. W. *Science* **195,** 260–262 (1977).

Identification of stem cells in small intestine and colon by marker gene *Lgr5*

Nick Barker[1], Johan H. van Es[1], Jeroen Kuipers[1], Pekka Kujala[2], Maaike van den Born[1], Miranda Cozijnsen[1], Andrea Haegebarth[1], Jeroen Korving[1], Harry Begthel[1], Peter J. Peters[2] & Hans Clevers[1]

The intestinal epithelium is the most rapidly self-renewing tissue in adult mammals. It is currently believed that four to six crypt stem cells reside at the +4 position immediately above the Paneth cells in the small intestine; colon stem cells remain undefined. *Lgr5* (leucine-rich-repeat-containing G-protein-coupled receptor 5, also known as *Gpr49*) was selected from a panel of intestinal Wnt target genes for its restricted crypt expression. Here, using two knock-in alleles, we reveal exclusive expression of *Lgr5* in cycling columnar cells at the crypt base. In addition, *Lgr5* was expressed in rare cells in several other tissues. Using an inducible Cre knock-in allele and the *Rosa26-lacZ* reporter strain, lineage-tracing experiments were performed in adult mice. The Lgr5-positive crypt base columnar cell generated all epithelial lineages over a 60-day period, suggesting that it represents the stem cell of the small intestine and colon. The expression pattern of *Lgr5* suggests that it marks stem cells in multiple adult tissues and cancers.

The absorptive epithelium of the small intestine is ordered into crypts and villi[1], which in the mouse turn over every three to five days. The massive rate of cell production in the crypts is balanced by apoptosis at the tips of the villi. Until now, intestinal stem cells have not been functionally identified owing to the lack of unique markers and the absence of stem-cell assays. The analysis of mouse chimaeras and mutagen-induced somatic clones[2,3] as well as the study of regeneration after injury have allowed an operational definition of stem-cell characteristics. Self-renewing stem cells cycle steadily to produce the rapidly proliferating transit-amplifying cells capable of differentiating towards all lineages. The estimated number of stem cells is between four and six per crypt[2]. Long-term DNA-label retention has tentatively located stem cells at 'position +4' directly above the Paneth cells[4]. Three differentiated cell types (enterocytes, goblet cells and enteroendocrine cells) form from transit-amplifying cells at the crypt–villus junction and continue their migration in coherent bands stretching along the crypt–villus axis. Each villus receives cells from multiple different crypts. The fourth principal differentiated cell type, the Paneth cell, resides at the crypt bottom. The colon epithelium contains crypts, but has a flat surface rather than carrying villi. This epithelium comprises two main differentiated cell types: the absorptive colonocytes and the goblet cells[1]. Until now, no stem cells have been identified in the colon.

Lgr5 is expressed in crypt base columnar cells

Because Wnt signals constitute the principal driving force behind the biology of the crypt[5], we hypothesized that some Wnt target genes may be specifically expressed in the stem cells. We have previously described the Wnt target gene programme in human colorectal cancer cells and found that it is physiologically expressed in intestinal crypts[6,7]. Of approximately 80 selected Wnt target genes[7], most were expressed in either Paneth cells or transit-amplifying cells. The *LGR5* gene, however, was expressed in a unique fashion. It behaved as a Wnt target gene, because its expression was extinguished on the induced inhibition of Wnt pathway activity by dominant-negative TCF4 in a cell system described earlier[6] (Fig. 1a, lane 1 versus 2). Accordingly,
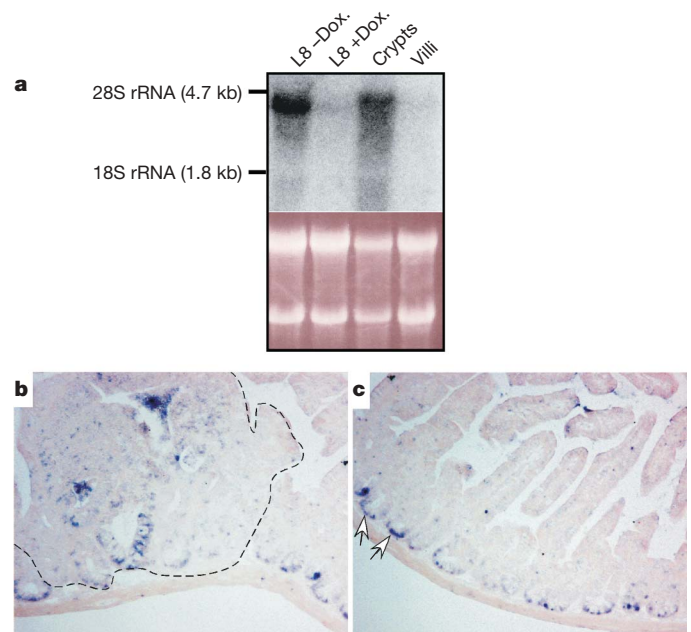

Figure 1 | *Lgr5* is a Wnt target gene in a human colon cancer cell line and is expressed in mouse crypts. a, Northern blot analysis (upper panel); ethidium-bromide-stained gel (lower panel). Lane 1: control Ls174T-L8 cells. Lane 2: Ls174T-L8 cells after 24-h doxycycline (Dox.)-induced Wnt pathway inhibition as in ref. 6. Note the strong downregulation of the 4.4-kilobase (kb) *Lgr5* messenger RNA on Wnt pathway inhibition. Lane 3: RNA extracted from isolated mouse small intestinal crypts, which unavoidably suffers from limited degradation resulting in some smearing. Lane 4: RNA extracted from isolated mouse villi. Note the specific expression of *Lgr5* in mouse crypts. **b, c,** Two overlapping images of an *in situ* hybridization performed on small intestines of an *Apc^Min* mouse, illustrating the ubiquitous expression of *Lgr5* at crypt bottoms (examples marked with white arrows) and the expression in the adenoma in **b** (marked by a dashed line).

[1]Hubrecht Institute, Uppsalalaan 8, 3584CT Utrecht, The Netherlands. [2]The Netherlands Cancer Institute, Antoni van Leeuwenhoek Hospital, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands.

the gene was expressed in the crypts, but not the villi, of mouse small intestine (Fig. 1a, lane 3 versus 4). *In situ* hybridization revealed expression in a limited number of cells located at all crypt bottoms as well as in adenomas in the small intestine of an *Apc*[Min] (a mutant allele of *Apc* as present in multiple intestinal neoplasia (*Min*) mice) mouse (Fig. 1b, c). This expression pattern, enlarged in Fig. 2c, clearly differed from that obtained with a Paneth-cell-specific gene (Fig. 2a) or a transit-amplifying-specific gene (Fig. 2b). The *Lgr5* gene seemed to mark the cycling crypt base columnar (CBC) cells, interspersed between Paneth cells (Fig. 2d–h).

Lgr5 encodes an orphan G-protein-coupled receptor, characterized by a large leucine-rich extracellular domain[8]. *Lgr5* was on our original list of Wnt targets in colorectal cancer[6], but has since been observed in ovarian and hepatocellular carcinomas[9,10]. To study its expression in detail, we obtained a knock-in allele, in which *lacZ*, preceded by an internal ribosome entry site (*IRES*), is integrated just amino-terminal to the first transmembrane domain essentially creating a null allele (Fig. 3a).

The *Lgr5*[−/−] phenotype was described in ref. 11. A malformation of the tongue and lower jaw causes newborn mutants to swallow air, leading to their death soon after birth. We observed the same phenotype. Of note, crypts and intestinal stem cells are first established several weeks after birth[12]. The generation of heterozygous *Lgr5-lacZ* mice allowed us to detail the expression of *Lgr5*. Before birth, a dynamic and complex expression pattern was observed (N. Barker *et al.*, manuscript in preparation). Around birth, *Lgr5* expression subsided in virtually all tissues. Expression in adult mice was restricted to rare, scattered cells in the eye, brain, hair follicle, mammary gland, reproductive organs, stomach and intestinal tract (Fig. 3, and not shown). In the small intestine, *Lgr5* expression was observed in slender cells at the bottom of small intestinal (Fig. 3b, c) and colon (Fig. 3d, e) crypts. Counting of blue cells in small-intestinal crypts
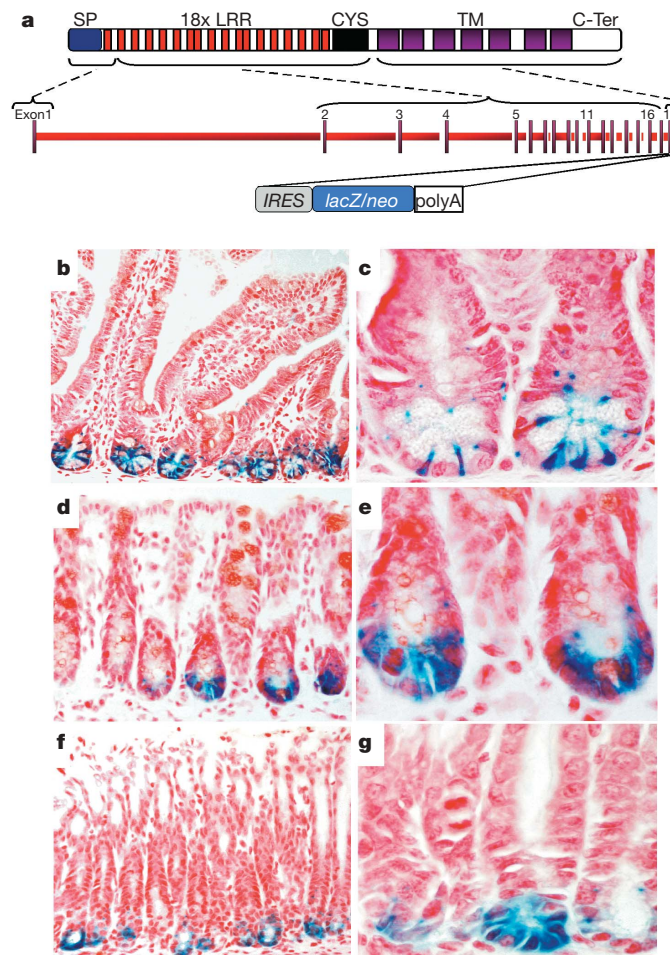

Figure 2 | *Lgr5* expression in cycling CBC cells of the small intestine. **a–c**, *In situ* hybridization was performed with probes specific for three Tcf target genes demonstrating non-overlapping expression patterns on the crypt epithelium. **a**, *Defa1* (also known as *Cryptdin1*) specifically marks Paneth cells at the crypt base; **b**, *Wdr43* (also known as *KIAA0007*) marks the transit-amplifying cells located above the Paneth cells; **c**, *Lgr5* is specifically expressed in 4–8 cells intermingled with the Paneth cells at the crypt base. All sense controls were negative (not shown). **d**, CBC cells (circled) are only poorly visible on haematoxylin and eosin (H&E)-stained sections. **e**, CBC cells (circled) are Ki67[+]. **f**, Some CBC cells express the M-phase marker phospho-histone H3 (circled). **g**, BrdU incorporation in CBC cells 4 h after a single dose of BrdU (circled). **h**, BrdU incorporation in CBC cells after 24 h continuous BrdU labelling (circled). **i**, Quantification of BrdU-labelled and LacZ-positive CBC cells per crypt. Four independent stretches of proximal small intestine totalling 400 crypts were counted. Results are depicted as means and standard deviations of numbers of positive CBC cells per crypt. The black bars show numbers of BrdU-positive CBC cells per crypt section after 4 h or 24 h; the white bar shows total number of CBC cells per crypt section assessed by counting LacZ-positive cells in *Lgr5-lacZ* mice.

Figure 3 | Restricted expression of an *Lgr5-lacZ* reporter gene in adult mice. **a**, Generation of mice carrying *lacZ* integrated into the last exon of the *Lgr5* gene, removing all transmembrane (TM) regions of the encoded Lgr5 protein. Neo, neomycin resistance cassette; SP, signal peptide; LRR, leucine-rich repeat region; C-Ter is carboxy terminus. **b–h**, Expression of Lgr5-LacZ (blue) in selected adult mouse tissues. **b, c**, In the small intestine, expression is restricted to six to eight slender cells intermingled with the Paneth cells at the crypt base. **d, e**, In the colon, expression is confined to a few cells located at the crypt base. **f, g**, Expression in the stomach is limited to the base of the glands.

sectioned through the lumen revealed the presence of approximately 3.5 of such cells per sectioned crypt (Fig. 2i, white bar). More than 30 yr ago, ref. 13 noted the presence of cycling cells between the Paneth cells; these were named 'crypt base columnar' cells. It has been proposed that these cells may harbour stem-cell activity[2,14,15].

Lgr5[+] CBC cells are actively cycling

By morphology, the slender Lgr5[+] CBC cells with their scant cytoplasm and wedge-shaped nuclei pointing towards the crypt lumen were readily distinguishable from the adjacent Paneth cells. The CBC cells were frequently positive for the proliferation marker Ki67 (Fig. 2e and Supplementary Fig. 1) and also occasionally expressed the M-phase marker phospho-histone H3, indicating that the cells are typically cycling (Fig. 2f). Indeed, a 4-h pulse of 5-bromodeoxyuridine (BrdU) labelled approximately one of these cells per crypt section (Fig. 2g and i, left black bar), whereas a 24-h continuous BrdU labelling resulted in more than three positive cells per crypt section (Fig. 2h and i, right black bar), close to the total number of CBC cells per crypt section (Fig. 2i, white bar). This observation implied that the average cycling time of CBC cells is in the order of one day.

Unique ultrastructural anatomy of Lgr5–EGFP[+] CBC cells

To visualize live CBC cells and to study their potential 'stemness', we generated another knock-in allele, by integrating an enhanced green fluorescent protein (EGFP)-IRES-creERT2 cassette at the first ATG codon (Fig. 4a and Supplementary Fig. 2). Heterozygous mice were healthy and fertile. The GFP pattern observed in adult tissues faithfully recapitulated the pattern previously seen with the Lgr5-lacZ allele (data not shown, and Fig. 4). Confocal imaging allowed the visualization of the Lgr5[+] cells by GFP fluorescence in small intestine (Fig. 4b, c, e) and colon (Fig. 4f). A three-dimensional reconstruction of the crypt in Fig. 4e (Supplementary Movie 1) illustrated the wedge-like shape of the CBC cells. Immuno-electron microscopy using immunogold labelling of the GFP-positive CBC cells and of neighbouring Paneth cells and fibroblasts illustrated the unique ultrastructural anatomy of the CBC cells (Fig. 4g, h). Typically, the CBC cells were relatively broad at their base, contained a flat, wedge-shaped nucleus and scarce organelles. A slender extension of apical cytoplasm was squeezed in-between neighbouring endoplasmic-reticulum- and granule-rich Paneth cells, extended to the crypt lumen and carried some apical microvilli.

We then crossed the EGFP-IRES-creERT2 knock-in allele with the Cre-activatable Rosa26-lacZ reporter[16] (Fig. 4a). Injection of tamoxifen activates the CreERT2 fusion enzyme in Lgr5-expressing cells. Cre-mediated excision of the roadblock sequence in the Rosa26-lacZ reporter should then irreversibly mark Lgr5[+] cells. Moreover, although potential progeny of these cells will no longer express GFP, the activated lacZ reporter should act as a genetic marker, facilitating lineage tracing.

Lgr5[+] CBC cells are distinct from the +4 cells

Expression of lacZ was not observed in non-induced mice (not shown). To visualize in the crypts the location of CBC cells in which the latent Cre enzyme could be activated by tamoxifen, we treated 2–3-month-old mice with tamoxifen and killed the mice 12-h later. As evident in Fig. 5a, blue LacZ signals appeared at the typical CBC positions. We determined the frequency at which the blue cells appeared at specific positions relative to the crypt bottom, according to the scheme in Fig. 5b. Most of the Cre[+], LacZ-labelled CBC cells occurred at positions between the Paneth cells, whereas only 10% of these cells were observed at the +4 position (Fig. 5b, blue line). Quantitative data on the position of long-term DNA-label-retaining cells obtained in adult mice after irradiation (marking the '+4' intestinal stem cell) were published recently[17]. Comparison of these data (Fig. 5b, red line) with the position of CBC cells harbouring

activatable Cre revealed that the two markers identified largely non-overlapping cell populations.

Another defining characteristic of the +4 cells is their exquisite sensitivity to low-dose (<1 Gy) radiation[4]. To compare relative radiation sensitivity between CBC cells and +4 cells, adult mice were irradiated with 1 Gy or 10 Gy and were killed 6 h later at the peak of apoptosis. Active caspase-3-positive cells were visualized by immunohistochemistry (Supplementary Fig. 3a). The frequency of caspase-3-positive cells per crypt was determined by counting
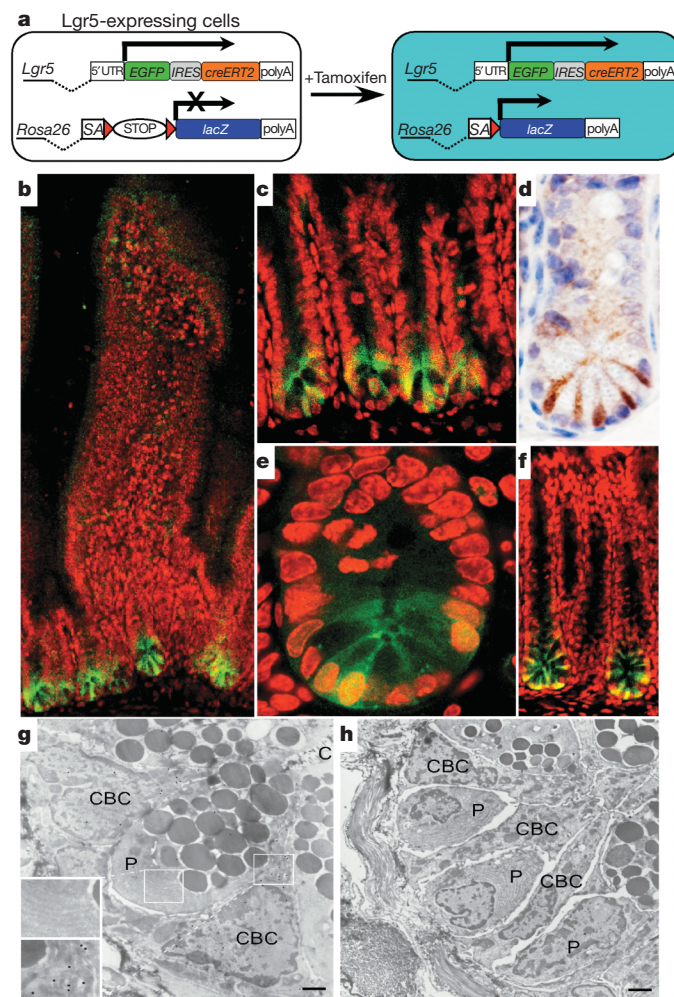

Figure 4 | EGFP expression in an Lgr5-EGFP-IRES-creERT2 knock-in mouse faithfully reproduces the Lgr5-lacZ expression pattern in the intestinal tract. a, Generation of mice expressing EGFP and creERT2 from a single bicistronic message by gene knock-in into the first exon of Lgr5. SA, splice acceptor; UTR, untranslated region. **b, c, e**, Confocal GFP imaging counter-stained with the red DNA dye ToPro-3 confirms that Lgr5 expression is restricted to the six to eight slender cells sandwiched between the Paneth cells at the crypt base of the small intestine. **b**, The entire crypt–villus unit; **c**, enlargement of crypt regions; **d**, immunohistochemical analysis of EGFP expression in intestinal crypts. **e**, Two-dimensional image of three-dimensional reconstruction (Supplementary Movie 1 and Supplementary Fig. 2). **f**, Confocal imaging of EGFP expression in the colon confirms that Lgr5 expression is restricted to a few cells located at the crypt base. **g**, Cryo-electron microscopy section of crypt stained for GFP with immunogold (scale bar, 1,000 nm). Quantification of specificity of labelling: gold particles were counted over 255 μm² of CBC cell cytosol (1,113 particles), 261 μm² of Paneth cell cytosol (305 particles) and 257 μm² of fibroblast cytosol (263 particles) outside the crypt. Thus, CBC cytoplasm had 4.36 gold particles per μm²; in contrast, the Paneth cells had 1.17 gold particles per μm² and the fibroblast control had 1.02 gold particles per μm². C, crypt lumen; P, Paneth cells. **h**, Unlabelled cryo-electron microscopy section (scale bar, 2,000 nm), underscoring the ultrastructural characteristics of CBC cells and their positions relative to Paneth cells.

1005

apoptotic cells in three classes: CBC cells (defined by their location between the Paneth cells), +4 cells (located directly above the Paneth cells) and transit-amplifying cells (located at position 5–15, Supplementary Fig. 3b). Maximal apoptosis at the +4 position was already reached at 1 Gy (Supplementary Fig. 3a, upper panel, black arrows) in concordance with ref. 4, whereas 10 Gy caused significantly more apoptosis than 1-Gy irradiation in CBC (Supplementary Fig. 3a, lower panel, white arrows) and transit-amplifying cells, confirming the different identities of the CBC and +4 cells.

Lgr5+ CBC cells are the small intestinal stem cells

Adult mice were then subjected to a tamoxifen pulse and were killed at 1 day (Fig. 5c), 5 days (Fig. 5d), 12 days, 35 days (not shown) and 60 days after induction (Fig. 5e and Supplementary Fig. 4). One day after induction, occasional CBC cells in small intestine and colon

Figure 5 | Lineage tracing in the small intestine and colon. a, *Lgr5-EGFP-IRES-creERT2* knock-in mouse crossed with *Rosa26-lacZ* reporter mice 12 h after tamoxifen injection. **b,** Frequency at which the blue cells appeared at specific positions relative to the crypt bottom, according to the scheme in the inset. Results are depicted as means and standard deviations of four independent stretches of proximal small intestine totalling 400 positive crypts. Most of the Cre+ LacZ-labelled CBC cells occurred at positions between the Paneth cells, whereas only 10% of these cells were observed at the +4 position directly above the cells (blue line). Quantitative data on the position of long-term DNA-label-retaining cells obtained in adult mice after irradiation (marking the '+4' intestinal stem cell) were published recently[17]. The graph shows a comparison of these data (red line) with the position of CBC cells carrying activated Cre. **c–e,** Histological analysis of LacZ activity in small intestine 1 day after induction (**c**), 5 days after induction (**d**) and 60 days after induction (**e**). **f–h,** Double-labelling of LacZ-stained intestine using PAS demonstrates the presence of goblet cells (**f,** white arrows) and Paneth cells (**g,** blue arrows) in induced blue clones. Double-labelling with synaptophysin demonstrates the presence of enteroendocrine cells within the induced blue clones (**h,** black arrows). **i–k,** Histological analysis of LacZ activity in colon 1 day after induction (**i**), 5 days after induction (**j**) and 60 days after induction (**k**).

expressed LacZ (Fig. 5c, i). As is demonstrated for whole-mount small intestine in Supplementary Fig. 4, parallel ribbons of cells emanated from the crypt bottoms and ran up the side of adjacent villi at later time points. For each of the time points, 200 crypts were counted in the proximal small intestine to determine the percentage containing blue cells. After 1 day, 5 days, 35 days and 60 days, these numbers were 22%, 39%, 25% and 36%, respectively. Thus, all CBC cells seemed capable of long-term maintenance of the self-renewing epithelium.

Double-labelling of 60-d-induced intestine revealed Periodic-Acid-Schiff (PAS)-positive goblet cells (Fig. 5f), PAS-positive Paneth cells (Fig. 5g) and synaptophysin-positive enteroendocrine cells (Fig. 5h) in the LacZ-stained clones. Using mutational marking, ref. 2 reported the existence of different types of long-lived epithelial clones; that is, columnar (enterocyte) clones, mucous (goblet) clones and mixed clones. Our clones were exclusively of the mixed variety. In blue clones, the frequency of goblet cells (114 of 2,043 total cells counted), enterocytes (1,846 of 2,043) and Paneth cells (83 of 2,043) was comparable to the frequency of goblet cells (127 of 3,691 total cells counted), enterocytes (3,345 of 3,691) and Paneth cells (127 of 3,691) in unmarked adjacent epithelium. As noted[2], the third secretory cell type, the enteroendocrine cell, was too rare to allow accurate enumeration. Taken together, we conclude that the Lgr5+ CBC cells represent the genuine stem cells of the small intestine.

Lgr5+ cells are the stem cells of the colon

Analysis of the colon yielded essentially identical observations. The Lgr5+ cells yielded blue clones emanating from the crypt bottom (Fig. 5i). These clones contained colonocytes as well as goblet cells, and essentially remained unchanged during the 60 days of chase (Fig. 5j, k). One important difference to the situation in the small intestine involved the kinetics of clone formation. At 5 days, blue staining in most crypts was still restricted to the bottom, and entirely blue crypts were only rarely observed, implying that the colon stem cells were more often quiescent than their small-intestinal counterparts. At later days, the relative number of entirely blue crypts increased. Lgr5+ colon cells fulfilled the stem-cell requirements in being pluripotent and capable of maintaining epithelial self-renewal over long periods of time.

Our observations provide the definitive characterization of the intestinal stem cell by lineage tracing using the expression of a single marker gene, *Lgr5*. The small intestinal Lgr5+ cells are generally not quiescent, but are rapidly cycling, as demonstrated by the expression of Ki67 and phospho-histone H3, by the incorporation of BrdU and by the kinetics of ribbon formation. Lgr5+ cells of the small intestine seem to divide more actively than their colonic counterparts, probably reflecting differences in the rate of epithelial turnover between the two organs. It seems counterintuitive that stem cells cycle. This is, however, not unprecedented. Germ stem cells in the *Drosophila* testis and ovary, arguably the best understood adult stem cells in animals, cycle throughout the lifetime of the adult fly[18]. Similarly, a recent elegant study demonstrated that adult stem cells of mammalian skin are continuously cycling[19].

The cycling +4 cells have previously been proposed to represent the small intestinal stem cells[4], a notion not confirmed here. This proposal was based on the observation that a DNA label incorporated during periods of high stem-cell activity was specifically retained in cells at the +4 position. Long-term label retention is often used as an indirect strategy to identify stem cells[12]. It should be noted, however, that terminally differentiating cells will also retain DNA labels, and that label retention should therefore be interpreted with caution. Previous studies have proposed other markers for intestinal stem cells. Musashi[20,21] and CD133 (ref. 22) in our hands stain up to 30–50 cells per crypt (not shown), which seem to encompass CBC cells as well as early transit-amplifying cells. Reference 23 described several molecular markers for the +4 cells, including phospho-PTEN, phospho-AKT and 14-3-3ζ. Our current study implies that the validity of these putative stem-cell markers should be reconsidered.

Lgr5 may mark stem cells in other adult tissues

It seems rather unique that adult stem cells can be identified on the basis of the expression of a single gene. This phenomenon may not be restricted to the intestine, because we observe highly restricted *Lgr5* expression in a variety of other tissues. Indeed, *Lgr5* was recently reported to be the second most highly upregulated gene as assessed by differential expression arraying on isolated hair follicle stem cells[24]. Moreover, preliminary lineage tracing experiments in the hair follicle support the notion that Lgr5[+] cells represent stem cells (N. Barker, H. Clevers and R. Toftgard, unpublished data). Whereas patterns of proliferation in stomach glands have indicated that the epithelial stem cells reside at the isthmus, halfway between the gland base and epithelial surface[25], we find *Lgr5* expressed at gland bottoms (Fig. 3f, g). Ongoing lineage tracing experiments imply that the entire glands derive from these cells (N. Barker and H. Clevers, unpublished data). In the mammary gland, stem cells reside in the basal epithelial layer[26], where we observe *Lgr5* expression (not shown). Lgr5 may thus represent a more general marker of adult stem cells. If true, the mouse models developed in the course of this study will allow the isolation as well as specific genetic modification of live adult stem cells in a variety of organs. We first identified *Lgr5* as a gene expressed in colon cancer cells[6]. It is expressed in other cancers[9,10] and, as described in the current study, is also in scattered cells in pre-malignant mouse adenomas. Future studies should explore the possibility that these Lgr5[+] malignant cells may represent cancer stem cells.

METHODS SUMMARY

Northern blotting and induced Wnt pathway inhibition in LS174T clone L8. This was performed as in ref. 6. Crypt and villus epithelial preparations for RNA isolation were generated from 1–2-cm lengths of intestine by 8 rounds of incubation in a solution containing phosphate-buffered-saline lacking Ca^{2+} and Mg^{2+} (PBS0), 1 mM EDTA and 1 mM EGTA at 4 °C for 10 min, followed by vigorous shaking. Fractions 3–4 and 7–8, comprising predominantly villi and crypts, respectively, were used for RNA isolation.

Mice. *Lgr5-lacZ* mice were generated by homologous recombination in embryonic stem cells targeting an *IRES-lacZ* cassette to the 5′ end of the last exon, essentially removing all transmembrane regions (Lexicon). *Lgr5-EGFP-IRES-creERT2* mice were generated by homologous recombination in embryonic stem cells targeting an *EGFP-IRES-creERT2* cassette to the ATG of *Lgr5*. *Rosa26-lacZ* Cre-reporter mice were obtained from the Jackson Laboratory.

Tamoxifen induction. Mice of >8 weeks were injected intraperitoneally with 200 μl tamoxifen in sunflower oil at 10 mg ml^{-1}.

BrdU injection. Mice were injected intraperitoneally at 4-h intervals with 200 μl BrdU solution in PBS at 5 mg ml^{-1}.

Tissue preparation for immunohistochemistry, *in situ* hybridization and LacZ analysis. This was performed as previously described[27]. *In situ* probes comprising a 1 kb N-terminal fragment of mouse *Lgr5* were generated from Image Clone 30873333. Antibodies: Ki67 antibody (Monosan), phospho-histone H3 (Campro Scientific), anti-synaptophysin (Dako) and anti-BrdU (Roche). Polyclonal rabbit anti-GFP was supplied by E. Cuppen.

Immuno-electron microscopy. Intestines were dissected and perfuse-fixed in 4% paraformaldehyde (PFA) in 0.2 M PHEM buffer (240 mM PIPES, 40 mM EGTA, 100 mM HEPES, 8 mM $MgCl_2$ pH 6.9), embedded in gelatine, cryosectioned with a Leica FCS cryoultratome and immunolabelled with polyclonal rabbit anti-GFP antibody. Samples were trimmed using a diamond Cryotrim 90 knife at −100 °C (Diatome), and ultrathin sections of 70 nm were cut at −120 °C using a Cryoimmuno knife (Diatome). For the low-magnification electron-microscope images, the 15-nm protein A-gold particles (UMCU) were briefly silver-enhanced with R-GENT SE-EM (Aurion)[28]. Aspecific binding was diminished by applying blocking solution (Aurion) before the primary antibody.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Gregorieff, A. & Clevers, H. Wnt signaling in the intestinal epithelium: from endoderm to cancer. *Genes Dev.* **19**, 877–890 (2005).
2. Bjerknes, M. & Cheng, H. Clonal analysis of mouse intestinal epithelial progenitors. *Gastroenterology* **116**, 7–14 (1999).
3. Winton, D. J. & Ponder, B. A. Stem-cell organization in mouse small intestine. *Proc. Biol. Sci.* **241**, 13–18 (1990).
4. Potten, C. S., Booth, C. & Pritchard, D. M. The intestinal epithelial stem cell: the mucosal governor. *Int. J. Exp. Pathol.* **78**, 219–243 (1997).
5. Korinek, V. *et al.* Depletion of epithelial stem-cell compartments in the small intestine of mice lacking Tcf-4. *Nature Genet.* **19**, 379–383 (1998).
6. van de Wetering, M. *et al.* The β-catenin/TCF-4 complex imposes a crypt progenitor phenotype on colorectal cancer cells. *Cell* **111**, 241–250 (2002).
7. Van der Flier, L. G. *et al.* The intestinal Wnt/TCF signature. *Gastroenterology* **132**, 628–632 (2007).
8. Hsu, S. Y., Liang, S. G. & Hsueh, A. J. Characterization of two LGR genes homologous to gonadotropin and thyrotropin receptors with extracellular leucine-rich repeats and a G protein-coupled, seven-transmembrane region. *Mol. Endocrinol.* **12**, 1830–1845 (1998).
9. McClanahan, T. *et al.* Identification of overexpression of orphan G protein-coupled receptor GPR49 in human colon and ovarian primary tumors. *Cancer Biol. Ther.* **5**, 419–426 (2006).
10. Yamamoto, Y. *et al.* Overexpression of orphan G-protein-coupled receptor, Gpr49, in human hepatocellular carcinomas with β-catenin mutations. *Hepatology* **37**, 528–533 (2003).
11. Morita, H. *et al.* Neonatal lethality of LGR5 null mice is associated with ankyloglossia and gastrointestinal distension. *Mol. Cell. Biol.* **24**, 9736–9743 (2004).
12. Reya, T. & Clevers, H. Wnt signalling in stem cells and cancer. *Nature* **434**, 843–850 (2005).
13. Cheng, H. & Leblond, C. P. Origin, differentiation and renewal of the four main epithelial cell types in the mouse small intestine. V. Unitarian Theory of the origin of the four epithelial cell types. *Am. J. Anat.* **141**, 537–561 (1974).
14. Bjerknes, M. & Cheng, H. The stem-cell zone of the small intestinal epithelium. III. Evidence from columnar, enteroendocrine, and mucous cells in the adult mouse. *Am. J. Anat.* **160**, 77–91 (1981).
15. Stappenbeck, T. S., Mills, J. C. & Gordon, J. I. Molecular features of adult mouse small intestinal epithelial progenitors. *Proc. Natl Acad. Sci. USA* **100**, 1004–1009 (2003).
16. Soriano, P. Generalized *lacZ* expression with the ROSA26 Cre reporter strain. *Nature Genet.* **21**, 70–71 (1999).
17. Potten, C. S., Owen, G. & Booth, D. Intestinal stem cells protect their genome by selective segregation of template DNA strands. *J. Cell Sci.* **115**, 2381–2388 (2002).
18. Ohlstein, B., Kai, T., Decotto, E. & Spradling, A. The stem cell niche: theme and variations. *Curr. Opin. Cell Biol.* **16**, 693–699 (2004).
19. Clayton, E. *et al.* A single type of progenitor cell maintains normal epidermis. *Nature* **446**, 185–189 (2007).
20. Nishimura, S., Wakabayashi, N., Toyoda, K., Kashima, K. & Mitsufuji, S. Expression of Musashi-1 in human normal colon crypt cells: a possible stem cell marker of human colon epithelium. *Dig. Dis. Sci.* **48**, 1523–1529 (2003).
21. Potten, C. S. *et al.* Identification of a putative intestinal stem cell and early lineage marker; musashi-1. *Differentiation* **71**, 28–41 (2003).
22. O'Brien, C. A., Pollett, A., Gallinger, S. & Dick, J. E. A human colon cancer cell capable of initiating tumour growth in immunodeficient mice. *Nature* **445**, 106–110 (2007).
23. He, X. C. *et al.* BMP signaling inhibits intestinal stem cell self-renewal through suppression of Wnt–β-catenin signaling. *Nature Genet.* **36**, 1117–1121 (2004).
24. Morris, R. J. *et al.* Capturing and profiling adult hair follicle stem cells. *Nature Biotechnol.* **22**, 411–417 (2004).
25. Bjerknes, M. & Cheng, H. Multipotential stem cells in adult mouse gastric epithelium. *Am. J. Physiol. Gastrointest. Liver Physiol.* **283**, G767–G777 (2002).
26. Sleeman, K. E. *et al.* Dissociation of estrogen receptor expression and *in vivo* stem cell activity in the mammary gland. *J. Cell Biol.* **176**, 19–26 (2007).
27. Muncan, V. *et al.* Rapid loss of intestinal crypts upon conditional deletion of the Wnt/Tcf-4 target gene c-Myc. *Mol. Cell. Biol.* **26**, 8418–8426 (2006).
28. Peters, P. J., Bos, E. & Griekspoor, A. in *Current Protocols in Cell Biology* 4.7.1–4.7.9 (Wiley, New York, 2006).

METHODS

Generation of *Lgr5-EGFP-IRES-creERT2* mice. The expression construct was generated by cloning the various components into the polylinker of pBluescript SK+ (Stratagene) as depicted in Supplementary Fig. 2. Lgr5 flanking arms were generated by high-fidelity polymerase chain reaction from 129S7-derived genomic BAC clones. All components were sequence-verified. The expression construct (100 μg) was linearized and transfected into male 129/Ola-derived IB10 embryonic stem cells (provided by The Netherlands Cancer Institute) by electroporation (800 V, 3 μF). Recombinant embryonic stem cell clones expressing the neomycin gene were selected in medium supplemented with G418 (gentamycin sulphate; the effective concentration of which was determined by its kill curve) over a period of 7 days. Approximately 500 recombinant embryonic stem cell clones were picked into duplicate 96-well tissue culture plates and cultured to 70% confluency. One plate was cryo-preserved and the other plate was used for DNA isolation according to ref. 29. DNA from embryonic stem cells was screened for the presence of homologous recombinants by Southern blotting according to the strategy outlined in Supplementary Fig. 2. Positive clones were thawed, expanded into 6-well tissue-culture plates and injected into C57BL/6 blastocysts using standard procedures. Male chimaeras born after transplantation of the bastocysts into C57BL/6 foster mothers were subsequently mated with C57BL/6 females, and germline transmission was confirmed by screening offspring for the presence of the expression cassette in the *Lgr5* locus by polymerase chain reaction. The neomycin expression cassette was then excised *in vivo* by crossing the mice with EIIa-Cre mice as previously described[30].

β-galactosidase (LacZ) staining protocol. Organs were isolated and immediately incubated for 2 h in a 20-fold volume of ice-cold fixative (1% formaldehyde, 0.2% gluteraldehyde and 0.02% NP40 in PBS0) at 4 °C on a rolling platform. Intestines, colon and stomach were first cleaned by flushing with fixative to remove faeces and undigested food. The fixative was removed and the tissues washed twice in PBS0 for 20 min at room temperature (20 °C) on a rolling platform. The β-galactosidase substrate (5 mM $K_3FE(CN)_6$, 5 mM $K_4Fe(CN)_6 \cdot 3H_2O$, 2 mM $MgCl_2$, 0.02% NP40, 0.1% sodium deoxycholate and 1 mg ml^{-1} X-gal in PBS0) was then added and the tissues incubated in the dark overnight at room temperature. The substrate was removed and the tissues washed twice in PBS0 for 20 min at room temperature on a rolling platform. The tissues were then fixed overnight in a 20-fold volume of 4% PFA in PBS0 at 4 °C in the dark on a rolling platform. The PFA was removed and the tissues washed twice in PBS0 for 20 min at room temperature on a rolling platform.

The stained tissues were transferred to tissue cassettes and paraffin blocks prepared using standard methods. Tissue sections (4 μM) were prepared and counterstained with neutral red.

Whole-mount analysis of LacZ staining as shown in Supplementary Fig. 4 was performed on 150-μM tissue sections prepared on a vibratome (HM650V, Microm).

Confocal analysis of EGFP expression. 150-μm sections were prepared from non-fixed intestines using a vibratome and were fixed for 5 min in 4% formaldehyde. After counterstaining with ToPro-3 (Molecular Probes) the sections were analysed for EGFP expression by confocal microscopy (Leica SP2 AOBS).

The observed emission spectrum was compared with reference spectra to confirm that the observed fluorescence was EGFP-derived.

***In situ* hybridization protocol.** The *in situ* probes used in this study correspond to expressed sequence tags obtained from the IMAGE consortium (Geneservice Ltd; ImaGenes GmbH). The GenBank accession and IMAGE numbers for these probes are as follow: mouse *Lgr5*, CN70148, IMAGE 30873333; Cryptdin-1, AA871421, IMAGE 1096215; and *KIAA0007*, BF158985, IMAGE 3982366. To ensure the specificity of the probes, we generated both sense and antisense probes by *in vitro* transcription using DIG RNA labelling mix (Roche) according to the manufacturer's instructions.

Intestines from normal or *Apc^Min* mice were flushed and fixed overnight in formalin. Samples were then dehydrated and embedded in paraffin, sectioned at 8 μM and processed for hybridization as described below. Sections were de-waxed, re-hydrated, treated with 0.2 M HCl, digested in proteinase K solution, post-fixed, treated in acetic anhydride solution and hybridized overnight for 24–48 h at 68 °C with various probes in 5× SSC (pH 4.5), 50% formamide, 2% blocking powder (Roche), 5 mM EDTA, 50 μg ml^{-1} yeast transfer RNA, 0.1% Tween 20, 0.05% CHAPS and 50 μg ml^{-1} heparin. Sections were then rinsed in 2× SSC and washed for 3 × 20 min at either 60 °C or 65 °C in 2× SSC per 50% formamide. After several rinses in Tris-buffered saline containing 0.05% Tween (TBST), sections were then blocked for 30 min in TBST containing 0.5% blocking powder (Roche). Sections were subsequently incubated in blocking solution overnight at 4 °C with alkaline phosphatase-conjugated anti-digoxigenin (1/2,000 dilution, Roche). After washing several times in TBST, the colour reaction was performed with NBT/BCIP solution. For image analysis, sections were temporally mounted in glycerol or permanently mounted after dehydration in Pertex.

Isolation of intestinal crypt and villus fractions. Intestines were washed in ice-cold PBS (Mg^{2+}/Ca^{2+}) and sliced longitudinally to expose the crypts and villi. The intestine was then cut into small pieces (1–2-cm long) and washed several times in ice-cold PBS (Mg^{2+}/Ca^{2+}) to remove contaminants such as faeces and hair. The intestine pieces were incubated in a solution containing PBS0 (lacking Mg^{2+}/Ca^{2+}), 1 mM EDTA and 1 mM EGTA at 4 °C for 10 min on a rotating platform. The PBS0, containing loose pieces of mesenchyme and intestine, was decanted, cold PBS0 was added, and the tube was shaken vigorously ten times. The cells shed into the PBS0 were collected and labelled as fraction 1. The intestine pieces were again incubated in solution containing fresh PBS0, 1 mM EDTA and 1 mM EGTA at 4 °C for 10 min on a rotating platform. The PBS0 was decanted, fresh PBS0 was added and the tube was shaken vigorously 10–15 times. The cells shed into the PBS0 were filtered through a 70 μM nylon cell strainer (BD Falcon) and the flow-through discarded. The tissue retained on the filter, comprising villi, was stored in PBS (Mg^{2+}/Ca^{2+}) on ice (fraction 2). The incubation, shaking and straining steps were repeated until eight fractions had been collected. Fractions 3–6 comprised pure villus tissue, whereas fractions 7 and 8 isolated as the flow-through from the cell strainer comprised pure crypt tissue.

29. Ramírez-Solis, R. *et al.* Genomic DNA microextraction: a method to screen numerous samples. *Anal. Biochem.* **201**, 331–335 (1992).
30. Lakso, M. *et al.* Efficient *in vivo* manipulation of mouse genomic sequences at the zygote stage. *Proc. Natl Acad. Sci. USA* **93**, 5860–5865 (1996).

ARTICLES

Auxin transport is sufficient to generate a maximum and gradient guiding root growth

Verônica A. Grieneisen[1], Jian Xu[2], Athanasius F. M. Marée[1], Paulien Hogeweg[1] & Ben Scheres[2]

The plant growth regulator auxin controls cell identity, cell division and cell expansion. Auxin efflux facilitators (PINs) are associated with auxin maxima in distal regions of both shoots and roots. Here we model diffusion and PIN-facilitated auxin transport in and across cells within a structured root layout. In our model, the stable accumulation of auxin in a distal maximum emerges from the auxin flux pattern. We have experimentally tested model predictions of robustness and self-organization. Our model explains pattern formation and morphogenesis at timescales from seconds to weeks, and can be understood by conceptualizing the root as an 'auxin capacitor'. A robust auxin gradient associated with the maximum, in combination with separable roles of auxin in cell division and cell expansion, is able to explain the formation, maintenance and growth of sharply bounded meristematic and elongation zones. Directional permeability and diffusion can fully account for stable auxin maxima and gradients that can instruct morphogenesis.

The multitude of concentration-dependent effects of externally supplied auxins on plant development has long fuelled the idea that auxins form gradients in plants and generate developmental outputs[1–3]. Auxin distributions at the cellular level can be inferred from the activity of auxin-responsive promoter elements. This led to the discovery that auxin-response maxima in roots, embryos and shoot-derived primordia require polar auxin transport, mediated by efflux-facilitating transmembrane proteins of the PIN family[4–7]. In shoots, auxin maxima predict sites of primordia outgrowth[8,9] and leaf venation[10]; the role of auxin transport has been addressed in several modelling studies[11–15]. In the *Arabidopsis* root, an auxin response maximum is required for cell-type specification in its close vicinity, but also for cell division and cell expansion at a distance[4,16]. This stable auxin maximum has been assumed to necessarily involve, in addition to polar auxin transport, regulated auxin synthesis and/or breakdown, sources and/or sinks[5], yet the essential mechanisms underlying its formation and maintenance remained unclear. Furthermore, auxin response maxima may not only reflect differential auxin concentrations but also local changes in auxin signal transduction components[17,18]. Here we provide a minimal mechanism for auxin-based pattern formation, given experimentally known PIN localizations. We develop and experimentally validate a mesoscopic model on the basis of the central assumption that PIN proteins have a critical role in polar auxin transport, which is supported by the dramatic defects of combinations of *pin* mutants[7,16]. The model reveals that the PIN-mediated flux pattern can explain where and why a root auxin maximum and associated auxin gradients arise as robust manifestations of the auxin transport system.

A modelling framework for polar auxin transport

Auxin diffuses freely within cells and the cell wall and permeates cell membranes[19–21]. Most previous simulation models at the tissue/organ scale have mathematically combined diffusion and permeability into one effective flux[22]. In contrast, by taking the spatial structure of cells into account[23] in our model, diffusion and permeability are dealt with independently, using realistic parameter values (for parameter analysis see Supplementary Methods) rather than describing cells as units containing only one auxin concentration value with limited possible fluxes

between units. At the next scale of organization, our model specifies individual cell sizes within a structured tissue layout that describes the entire root tip. It represents cells as extended regions on a lattice characterized by appropriately scaled sizes and shapes (Fig. 1), including surrounding cell walls. The interplay of phenomena that arise from the different scales (cell shape and neighbourhood within the tissue) is essential for understanding the emerging auxin distribution.

Auxin influx and efflux are treated differently according to the following biochemical considerations. Most extracellular auxin is protonated, and thus freely permeates into cells—a process which may additionally be facilitated by AUX/LAX family influx transporters[24]. In the cytoplasm, auxin is de-protonated such that free passage outwards is very slow; consequently, transport facilitators are necessary that locally permeate the membrane and facilitate auxin efflux. PIN proteins have cell-type-specific polar localizations in a manner that fully correlates with auxin transport routes as determined by auxin labelling, local auxin production and bioassays[16,25,26], although other directional transporters may have additional roles[27,28]. Hence, in our model we specifically focus on the localization of the PINs (with permeability $P_{e_{pin}}$). We also allow for significant influx (P_i) and little efflux ($P_{e_{bg}}$); the latter is mainly due to a background of ubiquitously expressed PINs in all cells.

Although the specific cell layout of the *Arabidopsis thaliana* root served as a guide for the model, we simplified this layout to capture the essential features of root architecture and PIN localization required for generating and maintaining an auxin maximum. Cells differ in size depending on their position in the meristematic (MZ) or elongation zone (EZ). Characteristic polar PIN localization is specified within four generalized tissue types: a central region represents 'vascular' tissue with basally located PINs; peripheral regions that express PIN predominantly at the apex correspond to the 'epidermal' tissues; between these tissues, a subclass of the vascular tissue—the single 'border' cell layer—has basal–lateral expression of PINs; and finally, the distal 'cap' region expresses PINs ubiquitously along the cell membrane (Fig. 1). In the epidermal cell layer, very low expression of lateral PINs is experimentally observed. Although lateral inward-localized PINs give no qualitative effect on the auxin distribution, they do generate quantitative effects (Supplementary Fig. 1). We thus introduce them when they become necessary for the dynamics of root

[1]Theoretical Biology and Bioinformatics, [2]Molecular Genetics, Department of Biology, Utrecht University, Padualaan 8, 3584 CH, Utrecht, The Netherlands.

growth. The results shown below are robust with respect to varying the number of cell files chosen for the layout (Supplementary Fig. 2). We first investigated critical features of the transport system in a static cellular system and subsequently analysed the full model, including growth, in which the formation of a meristematic and expansion zone arises as a self-organizing property given a pre-set response function for auxin-regulated cell division, growth and expansion.

A polar-auxin-transport-dependent maximum

Simulations with auxins provided from the uppermost vascular region reach a steady state with a strong concentration peak spanning several cells and with a maximum in cells immediately above the cap region (Fig. 2a, b), corresponding to the experimentally observed auxin response maximum in *Arabidopsis* quiescent centre cells. The model indicates that the appearance of the auxin maximum depends only on the general layout of PIN-mediated permeability, because specific cellular activities, such as localized decay of auxin levels or auxin influx regulation, are explicitly not included. Variation of permeability over a wide parameter range, which could reflect either different transport rates or different membrane densities of PINs, does not cause the maximum to disappear (Supplementary Fig. 3). The stable concentration peak is accompanied by large auxin fluxes through the tissues; like a reversed fountain, the central flow downwards connects to the upward flow in the epidermal tissue through a redistributing root cap (Fig. 2c). The auxin increase is due to reflux of the upward flow all along the meristem back into the central downward flow, which captures auxin within a flux-loop, causing an increase of the concentration maximum, until a steady state is reached (Fig. 2b). The overall PIN layout specifies the region of the maximum

at the junction of the flows, which is positioned centrally just above the cap region. The density of lateral PINs, both those in the border and epidermal files (yellow and blue cells in Fig. 1), are crucial for the reflux and are the most sensitive determinants for the auxin-storage capacity of the root and the magnitude of the maximum (Supplementary Figs 1 and 3). The formation of the root's auxin maximum goes beyond the
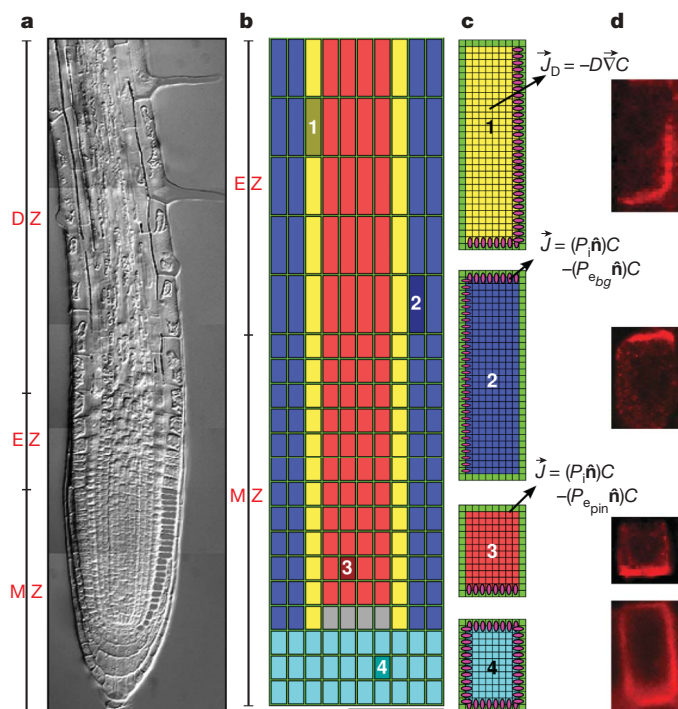

Figure 1 | Mesoscopic model for polar auxin transport. a, The *Arabidopsis* root. DZ, differentiation zone; EZ, elongation zone; MZ, meristematic zone. **b,** Generic root description adopted for the model. The four distinct regions treated in the simulations are indicated (vascular, red; epidermal layers, blue; border cells, yellow; and columella tiers, cyan). Quiescent centre (QC) cells (in grey) are equivalent to vascular cells. Scale bar, 100 μm. A cell wall/apoplast (green) surrounds all cells. Cells in the MZ and EZ have the same width, 16 μm, but differ in length: 60 μm (EZ) and 24 μm (MZ). **c,** PIN localization within the cells that compose these tissues, based on experimentally observed PIN distributions (**d**). For simplicity, influx carriers AUX (considered ubiquitously expressed in every cell) are not shown. For an explanation of the equations used, see Methods Summary and Supplementary Methods.
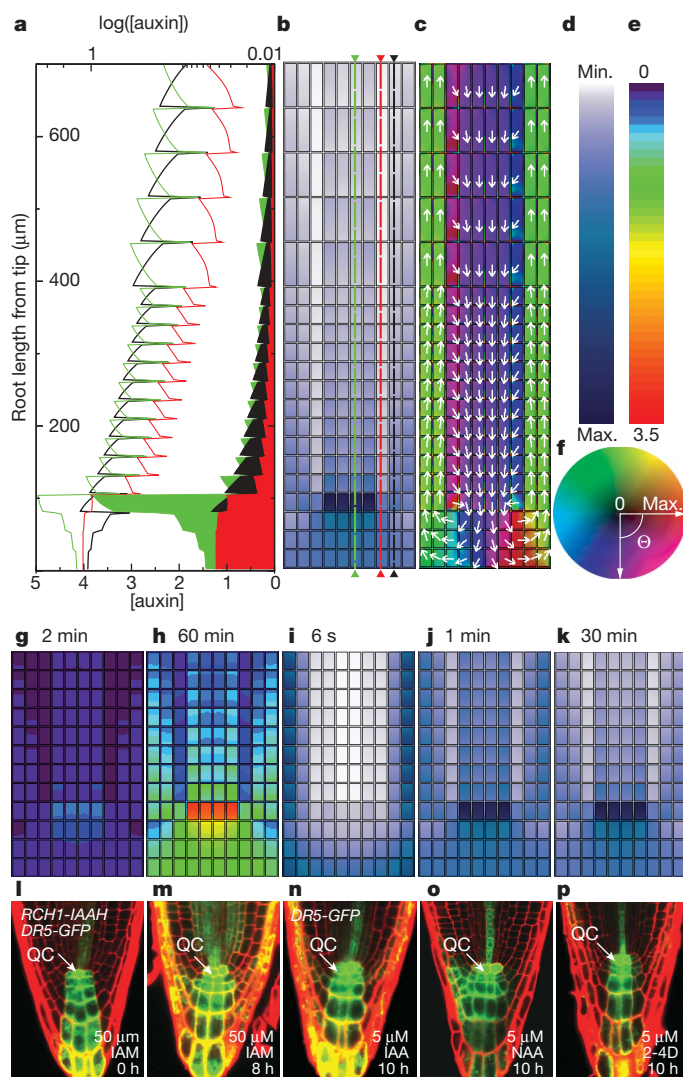

Figure 2 | Auxin maximum in simulations (a–k) and experiments (l–p). a, The steady-state concentration profiles along longitudinal sections through the different tissues as indicated by coloured arrows in **b**. Lines represent log([auxin]); filled profiles represent [auxin]; concentrations are in arbitrary units (A.U.), see Supplementary Methods 4.3. **b,** Steady-state concentrations within the root (Supplementary Movie 1), according to the colour bar (**d**) and scaled from the minimum (min.) to the maximum (max.) auxin concentration. **c,** Log of auxin fluxes measured in $\mu m^{-1} s^{-1}$ (A.U.) forming the flux–reflux pattern. The directionality and magnitude of the fluxes are given by the colour (arc) and intensity (radius) as indicated by the colour circle (**f**), as well as by the arrows. Ubiquitous endogenous auxin production, simulated (**g, h**) or experimentally induced (**l, m**), is shown. The established pattern after 1 min (**g**) and 60 min (**h**) (Supplementary Movie 2) is shown. The absolute concentration values are shown according to the colour bar (**e**). *DR5-GFP* expression in roots of *RCH1-IAAH* lines that are ubiquitously biosynthesizing auxin at 0 h (**l**) and 8 h (**m**) after induction by external addition of the precursor IAM (50 μM). A normal auxin distribution is observed. Root immersion in auxin, simulated (**i–k**) and in experiments (**n–p**), is shown. Establishment of the pattern, according to the colour bar (**d**), 6 s (**i**), 2 min (**j**) and 30 min (**k**) after immersion in auxin (Supplementary Movie 3, and Supplementary Methods for details). *DR5-GFP* expression in roots immersed in 5 μM IAA (**n**), NAA (**o**), and 2-4D (**p**), all after 10 h. All confocal images were taken using the same settings. In **g–k** we show the most distal 314 μm of the root.

classical description of uni-directional transport causing accumulation[29], because it essentially depends on global features of PIN localization throughout the root that allow for reflux (Supplementary Fig. 4). High auxin throughput (up to 30% of the intracellular auxin per second) generates a stable but highly dynamic auxin equilibrium profile that requires neither production nor decay in the roots.

Auxin equilibrates into the given profile solely as a result of the high auxin flows and PIN localizations. When all cells ubiquitously produce auxin, the same pattern is generated (Fig. 2g, h). Even when production is localized outside the auxin-maximum region, the pattern does not change (Supplementary Fig. 5). To test experimentally the prediction that the auxin maximum is insensitive to a ubiquitous rise in auxin levels, we stimulated the production of the natural auxin indoleacetic acid (IAA) in the root meristem using the meristem-specific *RCH1* promoter and the IAAH biosynthetic enzyme that can convert precursor IAM to IAA[16,30]. IAA levels were monitored with the semiquantitative and threshold dependent auxin response reporter *DR5-GFP*. As in the simulation, a persistent maximum is observed at the quiescent centre (Fig. 2l, m). We simulated an extreme case in which an auxin-free root lacking any influx from the shoot yet maintaining a normal efflux is immersed in an auxin solution. We find that auxin is polarly transported and rapidly redistributes in the immersed root tip, such that the auxin maximum is formed (Fig. 2i–k). Immersion of *Arabidopsis* roots in high concentrations of IAA indeed reveal the same persistent *DR5-GFP* pattern (Fig. 2n), further supporting the notion that the auxin maximum in the root tip is insensitive to the position of the auxin source. To show that the auxin transport system is robust to permeability rates, as predicted by our simulations, immersion experiments were performed with the synthetic auxins, 2,4-dichlorophenoxyacetic acid (2,4-D) and 1-naphthylacetic acid (NAA). The resulting auxin analogue distributions indicate the preservation of the maximum (Fig. 2o, p). The auxin transport mechanism maintains the auxin maximum as long as the throughput exceeds the rate of uptake.

Our model predicts autonomy of the root system from the shoot; the auxin maximum remains stable even when the shoot, that is, the auxin supply, is removed. Simulating a sudden cessation of shoot-derived auxin into the root, we find that, although total levels of auxin may slowly drop, the pattern around the maximum remains highly conserved (Fig. 3a, b). To test this prediction, we dissected roots from shoots, and monitored auxin levels in the root tip by expression of the auxin sensors *DR5-GFP* and *IAA2-GUS* (Fig. 3c, d). In accordance with the modelling result, persistent expression of different reporters in the maximum is observed for more than 10–30 days after root cut, corroborating that the auxin transport system can maintain the maximum (Fig. 3e–j). Importantly, the amount of reflux in the root determines the timescales of auxin-maxima maintenance. Thus, the root is analogous to a capacitor: it efficiently stores auxin, and it will, in the absence of shoot auxin supply and the presence of leakage and decay, maintain the maximum while only slowly undergoing auxin loss (Supplementary Fig. 6). The capacitance depends on the generalized layout of PIN localization in the root tissue. Only certain alterations in PIN distribution, which change the generalized layout cause drastic modifications in the auxin distributions (Supplementary Fig. 4).

Realistic cell sizes and diffusion and transport rates in our model allow us to investigate the timescale involved with the creation and re-establishment of an auxin maximum. In the simulations, the establishment of a pattern occurs very fast, dictated by the high throughput. Simulations of a quiescent centre ablation experiment by elimination of the cells containing the auxin maximum reveals a new maximum within the time range of minutes, which from then on steadily increases (Fig. 4a–c, upper panels). Even in a simulated root cut experiment, a new maximum is established rapidly after quiescent centre ablation (Supplementary Fig. 7). To validate the fast timescales of auxin-maximum formation by the reflux loop, we laser-ablated the quiescent centre of a dissected root. A shift in auxin concentrations occurs within 3 h as measured by *DR5-GFP*, which includes the GFP maturation time. Thus, a new maximum rapidly

arises from redistribution within the isolated root (Fig. 4a–c, lower panels). These experiments confirm that the timescale involved in establishing the typical auxin maximum is faster than the timescale of cell differentiation[31], and they show that the auxin maximum can be re-patterned without shoot auxin input.

To explore the mechanisms by which auxin-maximum positioning is taking place, we analysed early modifications in the auxin distribution seen in the *pin3;pin4;pin7* mutant by simulations and in the experimental root. In this triple mutant, we observe the auxin maximum shift towards the root cap together with a reduction in its intensity (Fig. 4e, lower panel). Generally, PIN activities in multiple *pin* mutant combinations are difficult to dissect owing to compensatory changes in gene expression. However, compensatory changes in PIN1 and PIN2 proteins in the *pin3;pin4;pin7* mutant can be visualized and do not occur at early stages (Supplementary Fig. 8). We simulated the triple mutant by inferring from the membrane localization that the lack of PIN3, PIN4 and PIN7 results in an extreme reduction of efflux permeability in the columella tiers, as well as a reduction of both basal-orientated permeabilities in the vascular region and lateral inwards permeabilities in the border cell file (see Fig. 4 legend). Our simulation results match with the experimentally observed basal dislocation of the auxin maximum and concentration drop (Fig. 4e, upper panel). We next questioned why these alterations take place. Experimentally, such an analysis is difficult because permeability manipulation of the separate regions is not trivial, owing to both distinct and overlapping expression domains of PINs 3, 4 and 7. However, in the model, we can explore independently the functional contribution of different regions. The new positioning of the maximum can be attributed to the significant reduction in efflux permeabilities within the columella cap, which now, instead of effectively redirecting the flow, accumulates auxin, dislocating the maximum distally (Fig. 4f). The decrease in magnitude of the dislocated maximum is not due to the reduction in vascular apical PIN expression, because the model shows that a
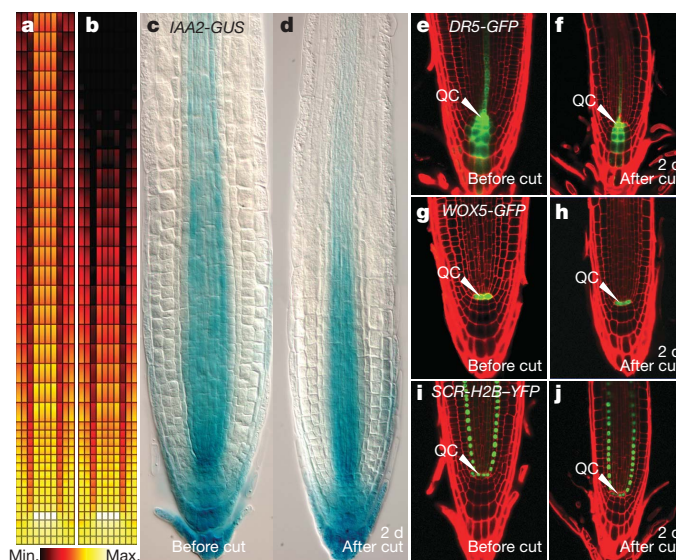

Figure 3 | An autonomous auxin capacitor in cut-root simulations (a, b) and experiments (c, d). a, Auxin distribution before cutting, that is, with auxin influx occurring at the uppermost vascular cells and efflux to the shoot occurring at the uppermost epidermal cells of a 1,700-μm-long root. **b,** Auxin distribution after cutting, that is, after alteration of the top boundary conditions such that there is no influx of auxin (Supplementary Movie 4). Colours indicate the relative log of auxin concentrations, according to the colour bar below. **c, d,** *IAA2-GUS* expression in a four-day-old intact root (**c**), and 2 d after the root has been cut (**d**), revealing after the cut a persistent auxin maximum and gradient in its vicinity, and a decrease in auxin levels in the EZ. Expression patterns of *DR5-GFP* (**e, f**), *WOX5-GFP* (**g, h**), and *SCR-H2B–YFP* (**i, j**), before (**e, g, i**) and 2 d after (**f, h, j**) root cut, revealing the correct maximum-associated gene-expression pattern.

reduction of this transport alone should generally lead to a build up of auxin concentrations in the vascular region (Fig. 4g). Instead, the concentration reduction is in fact due to the decrease in capacitance caused by the reduced lateral PIN density (Fig. 4h). Combined, these alterations yield the mutant pattern.

Our simulations consistently reveal a characteristic exponential basipetal auxin gradient associated with the auxin maximum (Fig. 2a). Modelling higher lateral permeabilities of the border cells and/or epidermal cells reveals that this should lead to a steeper gradient as well as a higher maximum (Supplementary Figs 1–3). Interestingly, our model predicts that the stabilization of the maximum-associated gradient takes much longer than the formation of the maximum itself (Supplementary Fig. 9). This difference in timescale reflects that the steady state, in which the total auxin levels remain constant, is dictated by a large capacity to store auxin. While the root system receives auxin input, it rapidly passes through a transient phase in which the auxin maximum is formed. Once the maximum is established, the root enters a quasi-steady state (QSS) in which auxin levels increase but the auxin distribution profile maintains its overall shape, characterized by an exponential gradient culminating at the maximum, and a proximal flat concentration profile given by the shoot influx. The distal region of the gradient is, on a logarithmic plot, seen as a linear slope that connects to the flat concentration region. As the accumulation of auxin proceeds, the QSS evolves by the expansion of the exponential part of

the profile into the proximal direction. In the absence of auxin decay, the system will practically always remain in the QSS, whereas with decay it remains in it for a period of days/weeks (dictated by the auxin half-life) before reaching equilibrium (Supplementary Fig. 9).

Emergence and dynamics of zonation in growing roots

We next considered the interplay between auxin dynamics and growth. We explore the pivotal issue of whether the auxin dynamics generated by our model produces developmentally instructive zonations (MZ, EZ) when simulating root growth. Auxin stimulates cell division and cell expansion at different concentrations in *in vitro* systems[32,33], and mutations in auxin response as well as in auxin-dependent patterning genes affect both cell division and cell expansion in roots[34,35]. Moreover, single or multiple *pin* mutants differentially affect cell division and cell expansion, further supporting spatially separable roles of auxin in the control of division and expansion zones[16]. As a simplifying approach we introduced a parsimonious description of the known auxin regulatory effects, that is, higher auxin levels promote cell division, whereas cell elongation is correlated with lower auxin levels. By implementing a graded response function that is identical for each cell and that dictates growth, division and rapid expansion (Fig. 5a, and Methods Summary for details), we explored whether growth feeds back on auxin distribution as well as if auxin concentrations alone could account for the formation of zonation. To capture possible slow developmental timescales, growth is simulated over an interval of 8 days.

At first observation, the growing root yields a similar auxin distribution to the static cell simulation, as long as the cellular growth rate does not yield a doubling time shorter than approximately 10–15 min, which is biologically unrealistic. The auxin maximum spans the same region during the growth process, maintaining itself in the correct position relative to the root tip (Fig. 5h, i). This robustness towards growth is due to the high auxin fluxes that accompany the maximum. Furthermore, our results reveal an emerging distal 'meristematic' region with small cells—and a proximal region where cell expansion prevails—caused by different levels of auxin within the self-generated auxin profile. Also in line with experimental data[36], mitotic activity in vascular regions extends further up (Fig. 5e) than in peripheral regions. This difference is solely caused by auxin profile concentrations that are slightly higher within the vascular region.

Interestingly, the low density of lateral localized PINs in the epidermal cell files plays an unexpectedly important part in the robustness of the developmental response to the auxin gradient during growth. A more spread out and uneven auxin distribution occurs in the absence of lateral epidermal PINs (the lower capacitance case) than in their presence (Fig. 5b, d, respectively), bringing forth developmental changes. Visualization of mitotic density distributions, that is, the number of mitoses that occur relative to the distance from the quiescent centre, reveals that mitotic regions are more defined in roots with higher capacitance (Fig. 5e) than in the lower-capacitance case (Fig. 5c). The lateral epidermal PINs assure a sharper transition between the MZ and EZ. This creates a small distinct region close to the root tip in which cells are small (MZ), with a sigmoidal transition towards more elongated cells (EZ). The lower capacitance root, lacking the lateral epidermal PINs, under the same parameters, does not reveal a clear distinct zonation between small and large cells (Fig. 5f).

The size of the meristematic zone slowly increases over the simulated 8 days, as can be seen in the cell-size distributions (Fig. 5g–i) and mitotic activities (Fig. 5j, k). This is due to the concentration profile being out of equilibrium (that is, in QSS), as can be inferred by the steadily increasing net auxin amount in the root system (inset of Fig. 5l). Although the auxin slope is preserved, the transition between the zones becomes sharper, as can be seen by comparing the mitosis profiles at 12 h and 8 days (Fig. 5j, k). The increasing sharpness is due to feedback between the dynamics of cell growth and the gradient: while the number of meristematic cells rises, the rate of root growth increases, thereby speeding up the dislocation of the accompanying
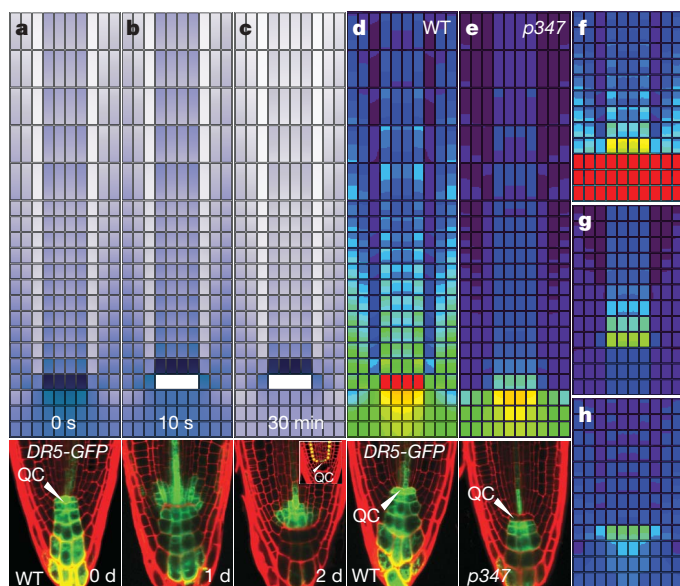

Figure 4 | Repositioning of the maximum owing to ablation and in the *pin* triple mutant, in simulations (a–e, upper panels, and f–h) and experiments (a–e, lower panels). **a–c**, Ablation; dynamics of new maximum establishment after ablation of the QC in a cut root; upper panels, steady state immediately before cut and ablation (**a**, upper panel) and 10 s (**b**, upper panel) and 30 min (**c**, upper panel) after cut and ablation (Supplementary Movie 5). The simulated root is 1,700-μm long, and here we show the most distal 700 μm. Colours are as shown in Fig. 1d. Lower panels, *DR5-GFP* expression before QC ablation (**a**, lower panel), 1 d after QC ablation (**b**, lower panel), and 2 d after QC ablation (**c**, lower panel). QC ablations were performed 8 h after the root was cut. The inset in the lower panel of **c** shows *pSCR-H2B–YFP* expression 2 d after QC ablation, suggesting that QC re-specification has occurred. **d–e**, Triple mutant *pin3;pin4;pin7* (*p347*) is simulated by reducing the permeability of the efflux values of vascular tissue and border cells to a quarter of their default values, and of columella tiers to a tenth of their default value. Default values and PIN orientations are as in Fig. 1 and Methods Summary. **d**, Wild-type auxin distribution. **e**, The resulting distribution in the triple mutant. Colours in upper panels of **d** and **e** give absolute concentration values according to the colour bar of Fig. 2e. **d**, Lower panel, *DR5-GFP* in WT. **e**, Lower panel, *DR5-GFP;p347*. **f–h**, Regional PIN contributions to the auxin distribution, with only root cap PIN reduction (**f**), only vascular PIN reduction (**g**), or only lateral PIN reduction of border cells (**h**).

auxin gradient. Consequently, a cell at the transition region experiences a more dramatic decline in concentration, which, combined with the cells' graded auxin response, causes a sharpening of the transition. The slow timescale growth process observed in the
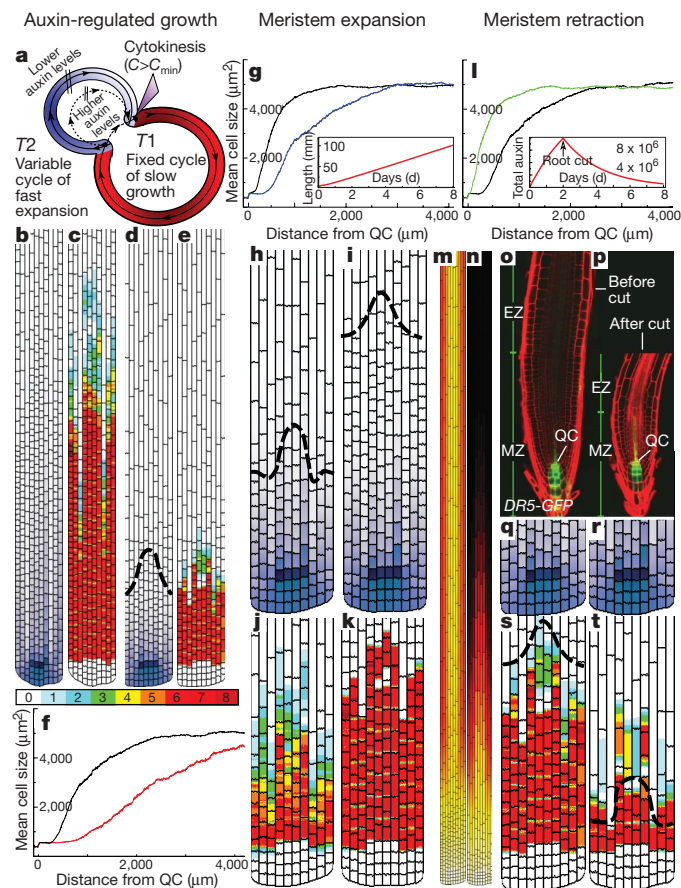

Figure 5 | Root growth and zonation in simulations (a–n, q–t) and experiments (o, p). a, Model for auxin-regulated cell growth and expansion (see Methods Summary and Supplementary Methods). **b–e,** Self-organization of the MZ and EZ; auxin distribution and meristematic distribution of cell divisions after 1 d of growth without (**b, c**) and with (**d, e**) lateral efflux permeability at the epidermal cell layers, determining lower- and higher-capacitance roots, respectively. Colours in **c** and **e** indicate the number of mitotic events over a 12 h interval that occurred during the growth at that position relative to the distance from the QC, with colour coding indicated below. **f,** Average cell size as a function of the distance from the QC at 1 d of growth, of both a lower- (without lateral PINs in the epidermis; red line) and a higher- (with lateral PINs in the epidermis; black line) capacitance root. **g,** Expansion of the MZ; spatiotemporal averages of cell size, showing the expansion of the meristematic zone between 12 h (black line) and 8 d (blue line). Inset shows total length grown with time revealing an almost constant, slowly increasing growth rate over 8 d. **h, i,** Auxin distribution and cell layouts at 12 h and 8 d of growth; **j, k,** the same MZ expansion seen through mitosis activity at the same time points (Supplementary Movie 6). **l,** Root cut and contraction of MZ; spatiotemporal cell size distributions at the moment of root cut (2 d; black line), and 6 d (green line) thereafter, showing meristem zone contraction after root cut. Inset reveals total auxin amount in a growing root receiving input from the shoot (first 2 days), and after root cut. **m, n,** Auxin distribution at the moment (**m**) and 6 d after (**n**) root cut, over the entire length of simulated root (Supplementary Movie 7). Colouring represents the log of auxin concentrations. **o, p,** DR5-GFP expression before (**o**) and 6 d after (**p**) root cut, showing conservation of the auxin maximum, and extension of the EZ into the MZ. **q, r,** Detailed images revealing auxin-maximum conservation: auxin distribution at the root tip at the moment of root cut (**q**) and 6 d thereafter (**r**). **s, t,** Mitosis activity at the moment of root cut (**s**), and 6 d thereafter (**t**), showing that meristematic activity becomes restricted nearer to the QC area after root cut. Dashed guiding lines in **d, h, i** and **s** help to visualize zonation that is due to cell size.

simulations is in accordance with the increase of MZ found in wild-type roots[37].

Simulating the root cut in the full model, yields a concentration drop in the upper region of the gradient (Fig. 5m, n), whereas typical distributions around the auxin maximum are not altered (Fig. 5q, r). After the root cut, the EZ/MZ boundary shifts (Fig. 5l), owing to the decay-driven auxin gradient shift (seen by a drop in total auxin shown in the inset of Fig. 5l). We tested this prediction by monitoring cell elongation in dissected experimental roots, and, indeed, recorded slow EZ expansion at the expense of the MZ (Fig. 5o, p). This effect was independently obtained by inhibiting shoot-derived polar auxin transport in intact seedlings using agarose blocks containing the polar auxin transport inhibitor NPA, again suggesting that the expansion of the EZ is caused by a reduction of total auxin levels (Supplementary Fig. 10). Taken together, our results show that EZ/MZ zonation is dynamically changing both in intact and dissected roots, whereas the auxin distribution around the quiescent centre is highly stable. Thus, the experimentally observed PIN distribution is able to explain fast, robust auxin-maximum formation, as well as the slow changes in zonation. Such seemingly uncorrelated processes thus do not necessarily depend on additional regulatory mechanisms.

Discussion

Our model strongly supports the hypothesis that plant roots are able to generate, through a combination of their topological structure, internal PIN distribution and basic properties such as diffusion and permeability, an auxin maximum and a highly robust gradient with morphogenic properties. This auxin distribution is robust at several levels: it preserves its characteristics through a broad set of permeability values, it is stable towards auxin production and decay, it is robust against drastic alterations in influx and efflux (such as root cut and tissue ablation), and it is not destabilized by cell division and expansion. The auxin maximum is formed without local auxin production or decay and without regulated influx. Thus, quiescent centre biosynthesis or 'sink' tissues are not essential for auxin-maximum localization. Lateral permeability enables roots to store and efficiently redistribute auxin. This has implications for interpreting experiments that track biosynthesis, because most endogenous production of auxin rapidly accumulates by auxin transport into the maximum, which does not necessarily correspond to the region of (highest) production. Moreover, the generalized PIN distribution used here is able to explain the location of the quiescent centre, as well as a characteristic exponential gradient of the auxin distribution. In addition, our mesoscopic modelling method reveals that information about the local neighbourhood exists, not only as an absolute concentration per cell, but also as a concentration gradient within the cell (Fig. 2a). This feature might be linked to auxin-dependent polar attributes of cells, such as the outgrowth of root hairs at the basal end of epidermal cells[38].

Our implementation of auxin-dependent cell division and cell expansion reveals that it is possible to establish functional zonations of plant organs using polar auxin transport only, as long as the root has a high-enough degree of 'capacitance', that is, sufficient reflux, to maintain a steep gradient of auxin. We consider it highly likely that auxin maxima in plants are associated with auxin gradients, even though technical limitations have hitherto prohibited direct visualization of such gradients. Indirect observations point to the existence of an auxin gradient. For example, the *PLETHORA* genes, which are expressed in conjunction with the root auxin maximum and require auxin responsive transcription factors for expression, influence distal cell fate, cell division and cell expansion[35]. Notably, PLT proteins form a functional concentration gradient that extends up to the elongation zone[39], which makes them good candidates for representing a read-out of a root auxin gradient.

The basic principles that endow roots with the capacity to form a developmentally robust and information-rich auxin gradient contrast with classical notions on developmentally instructive gradients,

which are generated by 'a source' at the location of the highest concentration with associated gradients attributed to diffusion and decay[29]. In our model, production and decay are not critical for auxin distributions as long as they occur on a slower timescale than the rapid auxin throughput. Moreover, it also differs from the classical concept of a transport-driven gradient, in which the slope results from the 'heaping up' of a morphogen at a 'dead end' (that is, a low throughput sink). Instead, the maximum and gradient are a consequence of the global PIN topology within the root, which establishes a reflux pattern with a high throughput of auxin. We have also shown that, after the establishment of the gradient, auxin throughput remains high, even (or rather, especially) at the location of accumulation. As a consequence of the reflux, the system is able to maintain global information during dynamic processes such as growth. Moreover, in contrast to the classical image of a frozen morphogenic gradient, here we show that a quasi-steady-state can transfer enough information to maintain a stem-cell-associated maximum, meristem and elongation zone, whereas, on a very different timescale, this profile changes, causing long-term developmental alterations. This reveals the importance of envisioning morphogenic gradients as non-static, out-of-equilibrium systems. In conclusion, transport-dependent dynamic auxin gradients constitute a powerful mechanism to generate developmental information.

METHODS SUMMARY

Static simulations of auxin transport (see Fig. 1). PINs and AUXs are located at the interfaces between cell and cell-wall sites, orientated opposite to one another ($\hat{\mathbf{n}}$ the unit vector normal to the surface of the membrane, orientated inwards), with permeabilities $P_{e_{pin}}$ and P_i of 20 μm s^{-1}. All cell membranes are given a small background efflux, $P_{e_{bg}}$, to account for membrane spreading of efflux carriers, which is set to 1 μm s^{-1}. The low epidermal lateral permeability $P_{e_{pin}}$ in the epidermis (represented as 'smaller' PINs in Fig. 1c, 2) is, for simplicity, set to zero in all static simulations, but used in the growing root case, in which it has a lower value of 5 μm s^{-1}. Auxin diffuses freely within cells and cell walls, with a diffusion coefficient of $D = 600$ μm^2 s^{-1}.

Simulations of growing roots (see Fig. 5a). The growth phase is described by a constant doubling time, $T1$; the expansion phase, with eightfold higher volume increases, is described by a variable duration, $T2$, inversely related to auxin level, C, at the centre of mass of each cell. $T2 = \max(\alpha \ln(C^*/C), 0)$, where C^* is the concentration above which cells only undergo the cycle $T1$. After $T1$ and $T2$, cytokinesis occurs if $C > C_{min}$. α is a sensitivity parameter. We include a cell-size check of minimal area 720 μm^2 for division, and a maximum cell area of 4,800 μm^2 above which cell expansion stops. Growth-related parameter values are derived from experimental data (Supplementary Material section 4.5) and set to: $T1 = 2$ h, $C_{min} = 3$, $C^* = 10$, $\alpha = 6.2$ h. The auxin decay rate is $\delta_{decay} = 5 \times 10^{-6}$ s^{-1}. Cell dynamics (growth and division) have been simulated by a modified Cellular Potts Model (see Supplementary Methods) in which turgor pressures drive cell movement within files. Cells expand in a longitudinal direction only (thus we assume stiff vertical cell walls); no constraints on relative cell file movements are included. Divisions are always anticlinal. Columella and quiescent centre cells don't divide. All simulations were run over the interval of 8 days, corresponding to 6,912,000 numerical time steps.

1. Berleth, T., Scarpella, E. & Prusinkiewicz, P. Towards the systems biology of auxin-transport-mediated patterning. *Trends Plant Sci.* **12**, 151–159 (2007).
2. Leyser, O. Auxin distribution and plant pattern formation: how many angels can dance on the point of PIN? *Cell* **121**, 819–822 (2005).
3. Jones, A. M. Auxin transport: down and out and up again. *Science* **282**, 2201–2203 (1998).
4. Sabatini, S. et al. An auxin-dependent distal organizer of pattern and polarity in the *Arabidopsis* root. *Cell* **99**, 463–472 (1999).
5. Friml, J. et al. AtPIN4 mediates sink-driven auxin gradients and root patterning in *Arabidopsis*. *Cell* **108**, 661–673 (2002).
6. Benková, E. et al. Local, efflux-dependent auxin gradients as a common module for plant organ formation. *Cell* **115**, 591–602 (2003).
7. Friml, J. et al. Efflux-dependent auxin gradients establish the apical–basal axis of *Arabidopsis*. *Nature* **426**, 147–153 (2003).
8. Reinhardt, D. et al. Regulation of phyllotaxis by polar auxin transport. *Nature* **426**, 255–260 (2003).
9. Heisler, M. G. et al. Patterns of auxin transport and gene expression during primordium development revealed by live imaging of the *Arabidopsis* inflorescence meristem. *Curr. Biol.* **15**, 1899–1911 (2005).
10. Scarpella, E., Marcos, D., Friml, J. & Berleth, T. Control of leaf vascular patterning by polar auxin transport. *Genes Dev.* **20**, 1015–1027 (2006).
11. Sachs, T. The control of the patterned differentiation of vascular tissues. *Adv. Bot. Res.* **9**, 152–262 (1981).
12. Feugier, F. G., Mochizuki, A. & Iwasa, Y. Self-organization of the vascular system in plant leaves: inter-dependent dynamics of auxin flux and carrier proteins. *J. Theor. Biol.* **236**, 366–375 (2005).
13. De Reuille, P. B. et al. Computer simulations reveal properties of the cell–cell signaling network at the shoot apex in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **103**, 1627–1632 (2006).
14. Jönsson, H., Heisler, M. G., Shapiro, B. E., Meyerowitz, E. M. & Mjolsness, E. An auxin-driven polarized transport model for phyllotaxis. *Proc. Natl Acad. Sci. USA* **103**, 1633–1638 (2006).
15. Smith, R. S. et al. A plausible model of phyllotaxis. *Proc. Natl Acad. Sci. USA* **103**, 1301–1306 (2006).
16. Blilou, I. et al. The PIN auxin efflux facilitator network controls growth and patterning in *Arabidopsis* roots. *Nature* **433**, 39–44 (2005).
17. Kepinski, S. & Leyser, O. The *Arabidopsis* F-box protein TIR1 is an auxin receptor. *Nature* **435**, 446–451 (2005).
18. Dharmasiri, N., Dharmasiri, S. & Estelle, M. The F-box protein TIR1 is an auxin receptor. *Nature* **435**, 441–445 (2005).
19. Gutknecht, J. & Walter, A. Transport of auxin (indoleacetic acid) through lipid bilayer membranes. *J. Membr. Biol.* **56**, 65–72 (1980).
20. Goldsmith, M. H. M., Goldsmith, T. H. & Martin, M. H. Mathematical analysis of the chemosmotic polar diffusion of auxin through plant tissues. *Proc. Natl Acad. Sci. USA* **78**, 976–980 (1981).
21. Kramer, E. M. & Bennett, M. J. Auxin transport: a field in flux. *Trends Plant Sci.* **11**, 382–386 (2006).
22. Kramer, E. M. Computer models of auxin transport: a review and commentary (2007). *J. Exp. Bot.* advance online publication doi:10.1093/jxb/erm060 (12 April 2007).
23. Kramer, E. M. PIN and AUX/LAX proteins: their role in auxin accumulation. *Trends Plant Sci.* **9**, 578–582 (2004).
24. Marchant, A. et al. AUX1 regulates root gravitropism in *Arabidopsis* by facilitating auxin uptake within root apical tissues. *EMBO J.* **18**, 2066–2073 (1999).
25. Rashotte, A. M., Poupart, J., Waddell, C. S. & Muday, G. K. Transport of the two natural auxins, indole-3-butyric acid and indole-3-acetic acid, in *Arabidopsis*. *Plant Physiol.* **133**, 761–772 (2003).
26. Wiśniewska, J. et al. Polar PIN localization directs auxin flow in plants. *Science* **312**, 883 (2006).
27. Geisler, M. et al. Cellular efflux of auxin catalyzed by the *Arabidopsis* MDR/PGP transporter AtPGP1. *Plant J.* **44**, 179–194 (2005).
28. Terasaka, K. et al. PGP4, an ATP binding cassette P-glycoprotein, catalyzes auxin transport in *Arabidopsis thaliana* roots. *Plant Cell* **17**, 2922–2939 (2005).
29. Wolpert, L. Positional information and the spatial pattern of cellular differentiation. *J. Theor. Biol.* **25**, 1–47 (1969).
30. Casamitjana-Martínez, E. et al. Root-specific *CLE19* overexpression and the *sol1/2* suppressors implicate a CLV-like pathway in the control of *Arabidopsis* root meristem maintenance. *Curr. Biol.* **13**, 1435–1441 (2003).
31. Xu, J. et al. A molecular framework for plant regeneration. *Science* **311**, 385–388 (2006).
32. Went, F. W. Growth, auxin, and tropisms in decapitated *Avena* coleoptiles. *Plant Physiol.* **34**, 236–249 (1942).
33. Skoog, F. & Miller, C. O. Chemical regulation of growth and organ formation in plant tissues cultured *in vitro*. *Symp. Soc. Exp. Biol.* **54**, 118–130 (1957).
34. Lincoln, C., Britton, J. H. & Estelle, M. Growth and development of the *axr1* mutants of *Arabidopsis*. *Plant Cell* **2**, 1071–1080 (1990).
35. Aida, M. et al. The *PLETHORA* genes mediate patterning of the *Arabidopsis* root stem cell niche. *Cell* **119**, 109–120 (2004).
36. Beeckman, T., Burssens, S. & Inzé, D. The peri-*cell*-cycle in *Arabidopsis*. *J. Exp. Bot.* **52**, 403–411 (2001).
37. Beemster, G. T. S. & Baskin, T. I. Analysis of cell division and elongation underlying the developmental acceleration of root growth in *Arabidopsis thaliana*. *Plant Physiol.* **116**, 1515–1526 (1998).
38. Masucci, J. D. & Schiefelbein, J. W. The *rhd6* mutation of *Arabidopsis thaliana* alters root-hair initiation through an auxin- and ethylene-associated process. *Plant Physiol.* **106**, 1335–1346 (1994).
39. Galinha, C. et al. PLETHORA proteins as dose-dependent master regulators of Arabidopsis root development. *Nature* doi:10.1038/nature06206 (this issue).

1013

ARTICLES

Kinetic redistribution of native and misfolded RNAs by a DEAD-box chaperone

Hari Bhaskaran[1] & Rick Russell[1]

DExD/H-box proteins are ubiquitously involved in RNA-mediated processes and use ATP to accelerate conformational changes in RNA. However, their mechanisms of action, and what determines which RNA species are targeted, are not well understood. Here we show that the DExD/H-box protein CYT-19, a general RNA chaperone, mediates ATP-dependent unfolding of both the native conformation and a long-lived misfolded conformation of a group I catalytic RNA with efficiencies that depend on the stabilities of the RNA species but not on specific structural features. CYT-19 then allows the RNA to refold, changing the distribution from equilibrium to kinetic control. Because misfolding is favoured kinetically, conditions that allow unfolding of the native RNA yield large increases in the population of misfolded species. Our results suggest that DExD/H-box proteins act with sufficient breadth and efficiency to allow structured RNAs to populate a wider range of conformations than would be present at equilibrium. Thus, RNAs may face selective pressure to stabilize their active conformations relative to inactive ones to avoid significant redistribution by DExD/H-box proteins. Conversely, RNAs whose functions depend on forming multiple conformations may rely on DExD/H-box proteins to increase the populations of less stable conformations, thereby increasing their overall efficiencies.

All organisms encode a host of RNAs that must fold into functional structures and undergo extensive conformational transitions as they mediate essential cellular processes, including translation and, in eukaryotes, pre-mRNA processing and the maintenance of chromosome ends. Essentially all processes that are mediated by structured RNAs also require DExD/H-box proteins, which use cycles of ATP binding and hydrolysis to accelerate conformational changes in RNA[1,2]. Although they are related to DNA helicases[3], and some viral DExD/H-box proteins possess at least modest canonical RNA or DNA helicase activity[4–6], many DExD/H-box proteins display very low activity in conventional helicase assays and are poorly processive, properties that are consistent with roles in disrupting local structural elements rather than unwinding long helices[2,7].

There has been much progress in using short model duplex RNAs and RNA–protein complexes to reveal the basic capabilities of DExD/H-box proteins[8–11]. However, relatively little is known at the molecular level about how DExD/H-box proteins interact with structured RNAs to mediate conformational changes, and what determines which RNAs, and which of their conformations, are targeted for action. Whereas many DExD/H-box proteins are thought to function in the context of a defined RNA or RNA–protein complex, some proteins of the largest subfamily, namely DEAD-box, function as general RNA chaperones by interacting less discriminately with structured RNAs to promote their folding[12,13]. This latter group includes the *Neurospora crassa* CYT-19 protein, which is required for the efficient splicing of several mitochondrial group I introns and can facilitate the folding of a diverse set of group I and group II introns *in vitro* or when expressed in *Saccharomyces cerevisiae*[12–14].

These systems are valuable experimentally because the relatively simple group I RNAs are likely to provide insight into the mechanisms of DExD/H-box proteins in more complex systems. A potentially useful candidate for detailed mechanistic studies of CYT-19 is the ribozyme derived from a group I intron of *Tetrahymena thermophila*, because its folding has been extensively characterized *in vitro*[15–19]. Further, because it folds preferentially to a long-lived

misfolded conformation that then slowly refolds to the native structure[20–24], it is possible to generate populations of either predominantly native or predominantly misfolded ribozyme. Using a quantitative assay for ribozyme catalytic activity, we showed previously that CYT-19 interacts with the misfolded ribozyme, giving ATP-dependent refolding to the catalytically active, native state[25,26]. Because the misfolded ribozyme is extensively structured, including all five long-range native tertiary contacts, and must unfold substantially to reach the native state[27], CYT-19 apparently accelerates this folding reaction by promoting partial unfolding of the misfolded ribozyme.

Redistribution of native and misfolded RNA

Here we probe the mechanism of CYT-19 action by exploring whether it recognizes structural features that are specific to the misfolded conformer or whether it can also mediate unfolding of the native ribozyme. We added CYT-19 and ATP to a preparation of ribozyme that was prefolded to the native state (Fig. 1a). At various times thereafter, we inactivated CYT-19 by adding proteinase K and increasing the Mg^{2+} concentration. We then determined the fraction of ribozyme present in the native conformation, without interference from CYT-19 (Supplementary Fig. 1), by measuring the fraction of added oligonucleotide substrate ($CCCUCUA_5$) that was rapidly cleaved by the ribozyme[25]. Our previous work showed that the substrate binds the native and misfolded species with similar rate constants but is cleaved only by the native ribozyme[22], so that a burst of product is obtained, with the amplitude indicating the fraction of native ribozyme (Fig. 1b). This burst is followed by a slower phase of product accumulation, which reflects dissociation of the substrate from the misfolded ribozyme and subsequent binding and cleavage by the native ribozyme[22,25].

As expected, no net unfolding was observed under conditions shown previously to give complete accumulation of native ribozyme (5 mM Mg^{2+}, Fig. 1c). However, under less stabilizing conditions

[1]Department of Chemistry and Biochemistry, Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, Texas 78712, USA.

(1 mM Mg^{2+}), the fraction of native ribozyme decreased on addition of high concentrations of CYT-19, indicating that CYT-19 can also unfold the native ribozyme (Fig. 1b–d; a subsequent slow increase reflects time-dependent inactivation of CYT-19 and inhibition by accumulating ADP (data not shown)). The reaction was dependent on ATP, because its omission or replacement with ADP or AMP-PNP (β-γ-imidoadenosine 5′-phosphate) gave only low levels of residual activity (Fig. 1c, d, and data not shown; higher concentrations of CYT-19 (not shown) gave significant ATP-independent activity, pre-sumably reflecting passive 'strand capture'[28]). The ribozyme ulti-mately reached an apparent steady state between the native and alternative conformers, which was dependent on CYT-19 concentra-tion (Fig. 1d). The observed rate constant for steady-state formation increased modestly with CYT-19 concentration, giving an efficiency of less than $10^5 M^{-1} min^{-1}$. The same steady-state distribution was obtained whether starting from a population of predominantly native or misfolded ribozyme (Fig. 1e), indicating that the entire ribozyme population is subject to the action of CYT-19.

Additional insight into the action of CYT-19 on the native ribo-zyme and the nature of the steady-state redistribution came from the finding that the decrease in native ribozyme was accompanied by formation of the previously characterized misfolded conformer. On inactivation of CYT-19 by proteinase K, the native ribozyme accumulated with the same rate constant as that for refolding from the long-lived misfolded species, and gave the same dependences on Mg^{2+} and urea concentrations (Fig. 1f and Supplementary Fig. 2), indicating that a large population of the misfolded conformer was either formed during the CYT-19 reaction or immediately on its inactivation. These observations suggested a simple model in which CYT-19 partly unfolds both the native and misfolded ribozyme spe-cies, giving intermediates that subsequently fold along the same path-way that predominates in the absence of CYT-19. This pathway includes a late commitment point for folding from a trapped inter-mediate (I_{trap}) to the native and misfolded species, with preferential partitioning to the misfolded species[18,22] (Scheme 1). This kinetic

preference could allow the misfolded species to accumulate despite having a lower stability than the native species.

Unfolding efficiency depends on RNA stability

To explore this model further, we were interested in identifying features of the ribozyme that affect the efficiency of unfolding by CYT-19. Although CYT-19 is able to act on the native ribozyme, the efficiency is estimated to be at least 50-fold lower than for the misfolded ribozyme (Fig. 1d and ref. 25). This difference raised the possibility that CYT-19 derives specificity from structural features of the misfolded species, but an alternative possibility was that the native ribozyme is unfolded less efficiently because it is more stable[22,29]. We therefore examined two ribozyme variants with decreased native stability (Fig. 2). First, we found that the native species of a ribozyme variant with a disrupted tertiary contact between the P4 helix and an A-rich internal loop in the peripheral helix P5a[27,30] was unfolded efficiently by CYT-19 even at 5 mM Mg^{2+} (Fig. 3a and Supplementary Fig. 3), conditions that do not give detectable unfolding of the wild-type ribozyme. Again, the reaction was dependent on ATP (Supplementary Fig. 4) and, for a given CYT-19 concentration, the same steady state was reached when starting from populations of predominantly native or misfolded ribozyme (Fig. 3b).

To test the model that CYT-19 produces intermediates that sub-sequently fold along the pathway that predominates in its absence, we compared the unfolding reactions with kinetic simulations (curves in Fig. 3b). The data were well described by using the experimental value for CYT-19-mediated unfolding of the native species and values for subsequent folding that were determined in the absence of CYT-19 (Supplementary Figs 5 and 6). Thus, CYT-19 apparently gives partial unfolding of the native and misfolded species and then allows the unfolded intermediates to fold again without significant interference in the latter process.

Next we examined a ribozyme variant with the entire P5abc peri-pheral element deleted ($E^{\Delta P5abc}$; Fig. 2)[31]. This ribozyme populates the native and misfolded species nearly equally at equilibrium[29], leading to the prediction that the two species would be unfolded with comparable efficiency. CYT-19 acted readily on the native ribozyme to decrease the fraction of native species ($1.4 \times 10^6 M^{-1} min^{-1}$; Fig. 3c, d). However, when starting from a population of misfolded ribozyme, net refolding was not observed; instead, a steady state was reached with the fraction of native ribozyme being approximately

Figure 1 | Unfolding of native and misfolded *Tetrahymena* ribozyme.
a, Reaction scheme. **b**, Cleavage of substrate after incubation with 1 mM Mg^{2+}, 2 μM CYT-19 and 2 mM Mg^{2+}ATP for 0.25 min (orange), 0.67 min (magenta), 1 min (cyan), 2.5 min (red), 9.5 (blue) or 22 min (dark green). Light green, no CYT-19. **c**, Native ribozyme unfolding (1 mM Mg^{2+}). CYT-19 was 1 μM (open downward triangles), 2 μM (solid coloured circles) or 3 μM without (filled upward triangles) or with (open upward triangles) 2 mM Mg^{2+}ATP. Coloured circles show burst amplitudes from the corresponding curves in **b**. Open circles, no CYT-19; filled black circles, 2 μM CYT-19, 2 mM Mg^{2+}ATP, 5 mM Mg^{2+}. **d**, Plot of rate constant (circles) and steady-state value (triangles) against CYT-19 concentration. **e**, Approach to steady state from native (circles) or misfolded (triangles) ribozyme with 1.2 μM (blue) or 2 μM (red) CYT-19. **f**, Refolding to the native state (triangles) after unfolding by CYT-19 (diamonds) and inactivation by proteinase K. Circles, no CYT-19.

equal to the small fraction that avoids the misfolded species and folds to the native state directly (Fig. 3c). This behaviour supports the model: if the native and misfolded species are rapidly unfolded with nearly equal efficiencies, even at modest CYT-19 concentrations the relative populations will be determined by the kinetic partitioning that occurs between them during folding. Nevertheless, the lack of native ribozyme accumulation prevented a determination of the efficiency of CYT-19 for unfolding the misfolded ribozyme.

We circumvented this limitation by including a small excess of the group I intron-binding protein CYT-18 (ref. 32), which binds the native $E^{\Delta P5abc}$ ribozyme tightly[33] but does not bind stably to the misfolded ribozyme under these conditions (A. Chadee and R.R., unpublished observations), such that it traps any newly formed native ribozyme. Net refolding of misfolded $E^{\Delta P5abc}$ ribozyme to the native state was indeed observed. After correction for the bias of the ribozyme towards misfolding again after CYT-19 action[34], this measurement gave an efficiency for unfolding of the misfolded ribozyme of $3.1 \times 10^{6}\,M^{-1}\,min^{-1}$ (Fig. 3d). This value is about double that for unfolding of the native $E^{\Delta P5abc}$ ribozyme, similar to the equilibrium value of 1.4 (ref. 29). Thus, the relative stabilities of the native and misfolded species seem to have a central function in determining their efficiencies of unfolding by CYT-19, although we cannot exclude the possibility that the region of the ribozyme near P5abc is the preferred site of action by CYT-19 and that it is the local stability of this region, rather than the global RNA stability, that governs CYT-19 efficiency. Again, kinetic simulation of the approaches to a steady-state mixture of native and misfolded ribozyme (Fig. 3c) gave good agreement with the data based on the experimental values for CYT-19-mediated unfolding of the native and misfolded species and previous experimental values for folding of the $E^{\Delta P5abc}$ ribozyme in the absence of CYT-19 (ref. 34).

Model for kinetic redistribution by CYT-19

All of the results are consistent with the model shown in Fig. 4. CYT-19 unfolds both the native and misfolded species, with efficiencies that depend on their relative stabilities, and then allows folding to proceed along the same pathway that dominates in its absence[22]. This generates an ATP-dependent steady state in which the misfolded species, and presumably earlier folding intermediates, are populated to much greater extents than at equilibrium because their formation is favoured kinetically during folding. It is intriguing that CYT-19 does not seem to affect subsequent folding of the ribozyme even though folding proceeds through intermediates that are, by definition, less stable than the native and misfolded species. Most probably, CYT-19 transiently unfolds these intermediates also, but the unfolding is not observed because the same intermediates are readily re-formed. Physical studies, probably at the single-molecule level, will be necessary to explore the intermediates that result directly from CYT-19 action.

It was proposed before the discovery of RNA chaperones that proteins might be required to accelerate the formation of the most stable conformations of RNA by facilitating transitions that require unfolding[35,36], and these general ideas were made more specific a decade later in a model in which CYT-19 actively unfolds misfolded group I RNAs, allowing them multiple chances to fold properly[37]. Whereas it was assumed for simplicity that the native species would be impervious to further chaperone action and would therefore accumulate, here we find that CYT-19 can have the opposite effect, decreasing the fraction of native ribozyme even though it is the most stable species. The mechanism of CYT-19 action is presumably the same as proposed earlier, but because it acts on both native and misfolded species it changes the distribution from equilibrium to kinetic control, allowing an increased population of intermediates that can form rapidly, even if they are less stable.

Implications for structured RNAs

Our results suggest that, *in vivo*, DExD/H-box proteins can act with sufficient breadth and efficiency to allow structured RNAs to
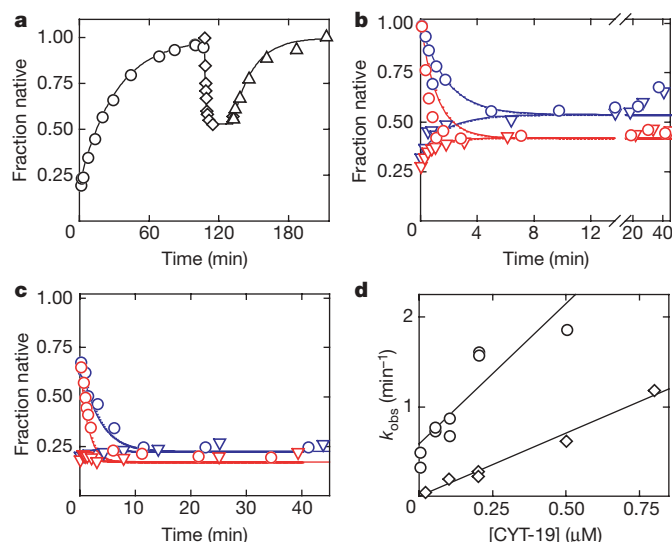

Figure 2 | Secondary structure, long-range tertiary contacts, and mutations of the *Tetrahymena* ribozyme. The five long-range tertiary contacts are indicated with red arrows. In the P5abc deletion variant ($E^{\Delta P5abc}$), the region shown in yellow is deleted, and nucleotides 126 and 196 are directly connected (thick grey line above yellow region). In the P5a mutant, nucleotides 183–188 (shaded cyan) are each changed to uridine[27]. This mutation disrupts the tertiary contact indicated by the black X.

Figure 3 | Unfolding of destabilized ribozyme variants. a, P5a variant with 0.8 μM CYT-19 (diamonds). On CYT-19 inactivation, native ribozyme accumulated (triangles; 0.056 min⁻¹) with the same rate constant within error as for refolding of the misfolded ribozyme (circles; 0.032 min⁻¹). **b**, Approach to steady state for native (circles) or misfolded (triangles) P5a variant with 0.5 μM (blue) or 1 μM (red) CYT-19. Curves depict kinetic simulations with experimentally derived values (see Supplementary Fig. 6), not fits to the data. **c**, Approach to steady state for $E^{\Delta P5abc}$ ribozyme. Curves and symbols are as in **b** except red (1.2 μM CYT-19). **d**, Unfolding of native (diamonds) and misfolded (circles) $E^{\Delta P5abc}$ ribozyme. All reactions were conducted at 25 °C, pH 7.0, 5 mM Mg²⁺.

populate a wider range of conformations than would be present at equilibrium. This redistribution of intermediates is analogous to an effect demonstrated for DEAD-box proteins by using model RNA duplexes of varying stability[11], and it is reasonable to imagine that depletion of the native state, rather than accumulation, is a general issue for RNA[38]. Indeed, simultaneous expression of an unrelated RNA chaperone protein and any of several mutant group I RNAs with decreased thermostability decreased their self-splicing activity *in vivo*, presumably because the chaperone unfolded the native RNAs[39]. The underlying reason for this behaviour is probably that RNA chaperones are unable to distinguish unambiguously between native and misfolded RNAs. Although an analogous challenge exists for protein chaperones, they can achieve a strong bias towards interacting productively with misfolded species by recognizing exposed hydrophobic surfaces[40]. In contrast, native and misfolded RNAs can be highly similar in global structure and perhaps identical on their surfaces[27].

Thus, RNAs that are required to populate one native structure may face selective pressure to minimize the extent to which their native structures are disrupted by DExD/H-box chaperones. Presumably, one important strategy is to 'hyper-stabilize' the native structures relative to alternative structures, beyond the level necessary simply to ensure their accumulation at equilibrium. Such an effect may contribute to the observation that the native state of the wild-type *Tetrahymena* ribozyme is about 6 kcal mol^{-1} more stable than the misfolded conformation under standard *in vitro* conditions[29], because this large energy gap results in the unfolding of the native species by CYT-19 so infrequently in comparison with the misfolded species that the kinetic preference for misfolding is overcome, and CYT-19 action yields an accumulation of the native state[25]. For many RNAs, proteins that associate to form functional complexes also presumably contribute to this energy gap.

There may be additional strategies available to structured RNAs to minimize the action of DExD/H-box proteins on their native states.
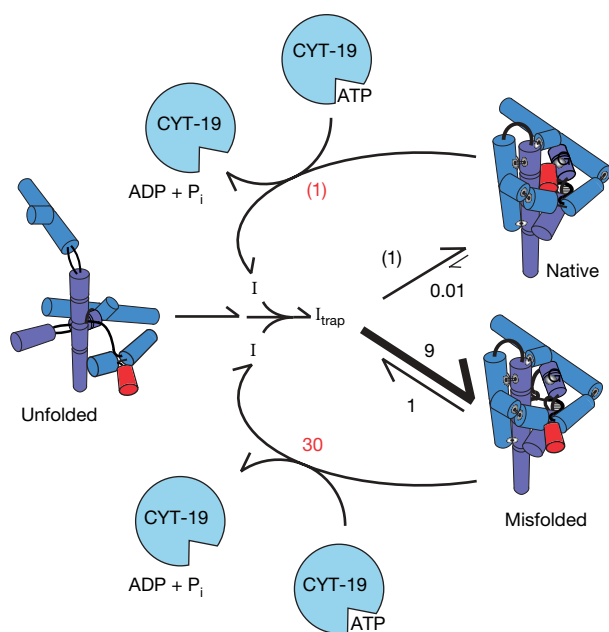
CYT-19 is strongly biased towards unwinding a duplex when it is unable to form tertiary contacts with the body of the *Tetrahymena* ribozyme[25], suggesting that tight packing of a native structured RNA would direct CYT-19 to interact more efficiently with extended or loosely packed misfolded structures. The quality control protein Ro selectively binds single-stranded ends of structured RNAs[41]; native RNAs can therefore presumably evade its action by protecting their ends. A further strategy, available for certain RNAs, is for the native structure to undergo a rapid and irreversible process, such as self-splicing, allowing mass action to drive the equilibrium towards the native form even if it is efficiently unfolded by chaperones.

Last, the ability of DExD/H-box proteins to increase the relative populations of less stable RNA structures is also likely to present opportunities for a broad range of RNAs whose functions require the formation of multiple structures. A striking example is the spliceosome, which relies on several DExD/H-box proteins to facilitate conformational changes during its reaction cycle[42,43]. Some of these transitions are likely to be thermodynamically unfavourable, as exemplified by the required separation of the extensively base-paired U4 and U6 small nuclear RNAs, and the action of DExD/H-box proteins may be necessary to prevent the more stable complex from dominating the steady-state population. MicroRNAs, which regulate gene expression by forming base-paired complexes with mRNA targets, may also use the action of chaperones to increase the sampling of alternative complexes, allowing a broader spectrum of physiological targets than would be expected from the relative stabilities of the target complexes[44]. This action of DExD/H-box proteins may also assist in the evolution of structured RNAs by allowing them to sample alternative structures, some of which could fortuitously possess beneficial activities[45] and would then be subject to further selection for stability and activity.

METHODS SUMMARY

Materials. Ribozymes were transcribed *in vitro* and purified with a Qiagen RNeasy column as described[25]. Oligonucleotide substrate (CCCUCUA$_5$; Dharmacon Research) was 5′ end-labelled with [γ-^{32}P]ATP by using T4 polynucleotide kinase and purified by non-denaturing PAGE[46]. CYT-19 was expressed and purified essentially as described[26].

Activity assay to monitor CYT-19-mediated unfolding. The fraction of native ribozyme was determined by monitoring substrate cleavage activity (Fig. 1a)[25]. For reactions monitoring unfolding of the native ribozyme, the ribozyme was prefolded to the native state in the presence of 10 mM Mg^{2+} (50 °C, 30 min). For reactions monitoring the unfolding of misfolded ribozyme and subsequent refolding to the native state, the ribozyme was incubated briefly with Mg^{2+} (5–10 min at 25 °C) to trap the misfolded species[22]. For all reactions, unfolding was initiated by the addition of CYT-19 in the presence of Mg^{2+}·ATP (2 mM unless indicated otherwise). At various times thereafter, aliquots were quenched for further CYT-19 action by adding 1 mg ml^{-1} Proteinase K and 50 mM MgCl$_2$ (ref. 25). To allow subsequent determination of the native fraction by activity, this quench solution included 500 μM guanosine, and for reactions of the E$^{\Delta P5abc}$ ribozyme it included 250 nM P5abc RNA (ref. 47). After incubation for a further 5–30 min, enough time for partly folded intermediates to complete folding predominantly to the misfolded state, but not long enough for significant refolding of misfolded ribozyme to the native state[22,27], the fraction of native ribozyme was determined by adding trace ^{32}P-labelled substrate and measuring the fraction cleaved rapidly in a burst (Fig. 1b)[22,27]. To determine the fraction of native ribozyme, burst amplitudes were divided by 0.80–0.85 to correct for a small fraction of inactive and presumably damaged ribozyme (15–20%; ref. 29).

Error estimates. Each experiment was performed at least twice. Variability between duplicates was typically 30–50% (Figs 1d and 3d, and Supplementary Fig. 3). In some cases, differences of up to threefold were observed between duplicate experiments performed with different preparations of CYT-19.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Figure 4 | Model for chaperone activity. CYT-19 generates ribozyme intermediates (I) and then allows them to fold again, thus increasing the population of less stable intermediates that are kinetically favoured. Values are for the P5a variant, normalized by native ribozyme unfolding (red; (1) indicates 5×10^5 M^{-1} min^{-1}; Supplementary Fig. 3), or for 'direct' native state formation (black; (1) indicates 0.4 min^{-1}; Supplementary Fig. 5). The wild-type ribozyme behaves similarly at low Mg^{2+} concentration, but at higher Mg^{2+} concentration the native species is sufficiently stable that CYT-19 unfolds it poorly and can therefore accelerate its formation from the kinetically trapped misfolded species.

1. Tanner, N. K. & Linder, P. DExD/H box RNA helicases: from generic motors to specific dissociation functions. *Mol. Cell* **8**, 251–262 (2001).

2. Jankowsky, E. & Fairman, M. E. RNA helicases—one fold for many functions. *Curr. Opin. Struct. Biol.* **17**, 316–324 (2007).

3. Gorbalenya, A. E. & Koonin, E. V. Helicases: amino acid sequence comparisons and structure–function relationships. *Curr. Opin. Struct. Biol.* **3**, 419–429 (1993).

4. Shuman, S. Vaccinia virus RNA helicase: an essential enzyme related to the DE-H family of RNA-dependent NTPases. *Proc. Natl Acad. Sci. USA* **89**, 10935–10939 (1992).

5. Jankowsky, E., Gross, C. H., Shuman, S. & Pyle, A. M. The DExH protein NPH-II is a processive and directional motor for unwinding RNA. *Nature* **403**, 447–451 (2000).

6. Pang, P. S., Jankowsky, E., Planet, P. J. & Pyle, A. M. The hepatitis C viral NS3 protein is a processive DNA helicase with cofactor enhanced RNA unwinding. *EMBO J.* **21**, 1168–1176 (2002).

7. Linder, P. Dead-box proteins: a family affair—active and passive players in RNP-remodeling. *Nucleic Acids Res.* **34**, 4168–4180 (2006).

8. Rogers, G. W. Jr, Richter, N. J. & Merrick, W. C. Biochemical and kinetic characterization of the RNA helicase activity of eukaryotic initiation factor 4A. *J. Biol. Chem.* **274**, 12236–12244 (1999).

9. Fairman, M. E. *et al.* Protein displacement by DExH/D RNA helicases without duplex unwinding. *Science* **304**, 730–734 (2004).

10. Yang, Q. & Jankowsky, E. The DEAD-box protein Ded1 unwinds RNA duplexes by a mode distinct from translocating helicases. *Nat. Struct. Mol. Biol.* **13**, 981–986 (2006).

11. Yang, Q., Fairman, M. E. & Jankowsky, E. DEAD-box-protein-assisted RNA structure conversion towards and against thermodynamic equilibrium values. *J. Mol. Biol.* **368**, 1087–1100 (2007).

12. Mohr, S., Stryker, J. M. & Lambowitz, A. M. A DEAD-box protein functions as an ATP-dependent RNA chaperone in group I intron splicing. *Cell* **109**, 769–779 (2002).

13. Huang, H. R. *et al.* The splicing of yeast mitochondrial group I and group II introns requires a DEAD-box protein with RNA chaperone function. *Proc. Natl Acad. Sci. USA* **102**, 163–168 (2005).

14. Mohr, S., Matsuura, M., Perlman, P. S. & Lambowitz, A. M. A DEAD-box protein alone promotes group II intron splicing and reverse splicing by acting as an RNA chaperone. *Proc. Natl Acad. Sci. USA* **103**, 3569–3574 (2006).

15. Latham, J. A. & Cech, T. R. Defining the inside and outside of a catalytic RNA molecule. *Science* **245**, 276–282 (1989).

16. Zarrinkar, P. P. & Williamson, J. R. Kinetic intermediates in RNA folding. *Science* **265**, 918–924 (1994).

17. Sclavi, B., Sullivan, M., Chance, M. R., Brenowitz, M. & Woodson, S. A. RNA folding at millisecond intervals by synchrotron hydroxyl radical footprinting. *Science* **279**, 1940–1943 (1998).

18. Treiber, D. K., Rook, M. S., Zarrinkar, P. P. & Williamson, J. R. Kinetic intermediates trapped by native interactions in RNA folding. *Science* **279**, 1943–1946 (1998).

19. Russell, R., Millett, I. S., Doniach, S. & Herschlag, D. Small angle X-ray scattering reveals a compact intermediate in RNA folding. *Nature Struct. Biol.* **7**, 367–370 (2000).

20. Russell, R. & Herschlag, D. New pathways in folding of the *Tetrahymena* group I RNA enzyme. *J. Mol. Biol.* **291**, 1155–1167 (1999).

21. Pan, J., Deras, M. L. & Woodson, S. A. Fast folding of a ribozyme by stabilizing core interactions: Evidence for multiple folding pathways in RNA. *J. Mol. Biol.* **296**, 133–144 (2000).

22. Russell, R. & Herschlag, D. Probing the folding landscape of the *Tetrahymena* ribozyme: Commitment to form the native conformation is late in the folding pathway. *J. Mol. Biol.* **308**, 839–851 (2001).

23. Treiber, D. K. & Williamson, J. R. Concerted kinetic folding of a multidomain ribozyme with a disrupted loop-receptor interaction. *J. Mol. Biol.* **305**, 11–21 (2001).

24. Russell, R. *et al.* Exploring the folding landscape of a structured RNA. *Proc. Natl Acad. Sci. USA* **99**, 155–160 (2002).

25. Tijerina, P., Bhaskaran, H. & Russell, R. Non-specific binding to structured RNA and preferential unwinding of an exposed helix by the CYT-19 protein, a DEAD-box RNA chaperone. *Proc. Natl Acad. Sci. USA* **103**, 16698–16703 (2006).

26. Grohman, J. K. *et al.* Probing the mechanisms of DEAD-box proteins as general RNA chaperones: The C-terminal domain of CYT-19 mediates general recognition of RNA. *Biochemistry* **46**, 3013–3022 (2007).

27. Russell, R. *et al.* The paradoxical behavior of a highly structured misfolded intermediate in RNA folding. *J. Mol. Biol.* **363**, 531–544 (2006).

28. Del Campo, M. *et al.* Do DEAD-box proteins promote group II intron splicing without unwinding RNA? *Mol. Cell.* (in the press).

29. Johnson, T. H., Tijerina, P., Chadee, A. B., Herschlag, D. & Russell, R. Structural specificity conferred by a group I RNA peripheral element. *Proc. Natl Acad. Sci. USA* **102**, 10176–10181 (2005).

30. Battle, D. J. & Doudna, J. A. Specificity of RNA-RNA helix recognition. *Proc. Natl Acad. Sci. USA* **99**, 11676–11681 (2002).

31. Joyce, G. F., van der Horst, G. & Inoue, T. Catalytic activity is retained in the *Tetrahymena* group I intron despite removal of the large extension of element P5. *Nucleic Acids Res.* **17**, 7879–7889 (1989).

32. Lambowitz, A. M., Caprara, M. G., Zimmerly, S. & Perlman, P. S. in *The RNA World* (eds Gesteland, R. F., Cech, T. R. & Atkins, J. F.) 451–485 (Cold Spring Harbor Laboratory Press, New York, 1999).

33. Mohr, G., Caprara, M. G., Guo, Q. & Lambowitz, A. M. A tyrosyl-tRNA synthetase can function similarly to an RNA structure in the *Tetrahymena* ribozyme. *Nature* **370**, 147–150 (1994).

34. Russell, R., Tijerina, P., Chadee, A. B. & Bhaskaran, H. Deletion of the P5abc peripheral element accelerates early and late folding steps of the *Tetrahymena* group I ribozyme. *Biochemistry* **46**, 4951–4961 (2007).

35. Karpel, R. L., Miller, N. S. & Fresco, J. R. Mechanistic studies of ribonucleic acid renaturation by a helix-destabilizing protein. *Biochemistry* **21**, 2102–2108 (1982).

36. Herschlag, D. RNA chaperones and the RNA folding problem. *J. Biol. Chem.* **270**, 20871–20874 (1995).

37. Thirumalai, D. & Hyeon, C. RNA and protein folding: common themes and variations. *Biochemistry* **44**, 4957–4970 (2005).

38. Mahen, E. M., Harger, J. W., Calderon, E. M. & Fedor, M. J. Kinetics and thermodynamics make different contributions to RNA folding *in vitro* and in yeast. *Mol. Cell* **19**, 27–37 (2005).

39. Grossberger, R. *et al.* Influence of RNA structural stability on the RNA chaperone activity of the *Escherichia coli* protein StpA. *Nucleic Acids Res.* **33**, 2280–2289 (2005).

40. Lin, Z. & Rye, H. S. GroEL-mediated protein folding: making the impossible, possible. *Crit. Rev. Biochem. Mol. Biol.* **41**, 211–239 (2006).

41. Stein, A. J., Fuchs, G., Fu, C., Wolin, S. L. & Reinisch, K. M. Structural insights into RNA quality control: the Ro autoantigen binds misfolded RNAs via its central cavity. *Cell* **121**, 529–539 (2005).

42. Staley, J. P. & Guthrie, C. Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell* **92**, 315–326 (1998).

43. Schwer, B. A new twist on RNA helicases: DExH/D box proteins as RNPases. *Nature Struct. Biol.* **8**, 113–116 (2001).

44. Rajewsky, N. microRNA target predictions in animals. *Nature Genet.* **38**, S8–S13 (2006).

45. Schultes, E. A. & Bartel, D. P. One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science* **289**, 448–452 (2000).

46. Zaug, A. J., Grosshans, C. A. & Cech, T. R. Sequence-specific endoribonuclease activity of the *Tetrahymena* ribozyme: enhanced cleavage of certain oligonucleotide substrates that form mismatched ribozyme–substrate complexes. *Biochemistry* **27**, 8924–8931 (1988).

47. van der Horst, G., Christian, A. & Inoue, T. Reconstitution of a group I intron self-splicing reaction with an activator RNA. *Proc. Natl Acad. Sci. USA* **88**, 184–188 (1991).

Author Contributions H.B. performed the experiments; R.R. and H.B. designed the study, interpreted results, and wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to R.R. (rick_russell@mail.utexas.edu).

METHODS

Purification of CYT-19. CYT-19 was purified as described[26] or with minor modifications. In the modified protocol, CYT-19 was expressed from *Escherichia coli* strain BL21 as a fusion with the maltose-binding protein (MBP). Induction was performed at 20 °C for 16 h in the presence of 1 mg ml^{-1} isopropyl-β-D-thiogalactopyranoside (IPTG). Cells were harvested by centrifugation, resuspended in 20 mM Tris-HCl pH 7.5, 500 mM KCl, 1 mM EDTA, and 2 mM dithiothreitol (DTT), and lysed by incubation for 20 min on ice in the presence of 1 mg ml^{-1} lysozyme followed by three sonication bursts of 15 s each (power setting 6 with a double-stepped microtip; Branson Sonifier S-450-A; VWR Scientific). The lysate was cleared by centrifugation (20,000*g* for 30 min at 4 °C), polyethyleneimine (PEI) was added slowly to 0.4%, and precipitated material was removed by centrifugation (18,500*g* for 15 min at 4 °C). The supernatant was loaded onto a 5-ml amylose column (high-flow resin; New England Biolabs) and washed with 3 column volumes of buffer containing 20 mM Tris-HCl pH 7.5, 500 mM KCl, 1 mM EDTA, 2 mM DTT. The column was then washed with 10 column volumes of the same buffer containing 1.5 M KCl and washed again with 10 column volumes of the same buffer containing 500 mM KCl. The fusion protein was eluted with the same buffer containing 10 mM maltose. Peak fractions were pooled (5 ml total) and incubated overnight at 4 °C in the presence of 40 µg ml^{-1} tobacco etch virus (TEV) protease to cleave the MBP tag. To remove the cleaved tag, the preparation was loaded on a 2-ml ceramic hydroxyapatite column (CHT; Bio-Rad Laboratories). This step differed from the earlier protocol[26], which used a gravity-flow hydroxyapatite column followed by a second application to the amylose column. The CHT column was washed with 40 column volumes of 20 mM potassium phosphate pH 7.3, 200 mM NaCl; CYT-19 was eluted with 500 mM potassium phosphate pH 7.3. Peak fractions were pooled (3 ml), dialysed overnight at 4 °C against 50 volumes of storage buffer (20 mM Tris-HCl pH 8.5, 500 mM KCl, 1 mM EDTA, 0.2 mM DTT, 50% (v/v) glycerol) and stored at −80 °C.

Activity assay to monitor CYT-19-mediated unfolding. The ribozyme was first incubated to give the desired starting conformation. Wild-type and variant ribozymes (2 µM) were incubated in the presence of 10 mM Mg^{2+} to yield predominantly the native conformation (50 °C for 30 min)[48] or predominantly the misfolded conformation (25 °C for 10 min, yielding about 90% misfolded ribozyme and 10% native ribozyme[22]). For each reaction, the starting fraction of native ribozyme was determined by adding trace ^{32}P-labelled substrate to an aliquot and processing it as described below. These reactions confirmed that the incubations yielded populations of predominantly native or misfolded ribozyme as expected. Incubation of the E^{ΔP5abc} ribozyme at 50 °C gave only about 60% native ribozyme, with the remainder misfolded, also as expected from previous work[29].

To initiate unfolding, CYT-19 was added at 25 °C. For control reactions in the absence of CYT-19, an equivalent volume of the storage buffer for CYT-19 was added so that all reactions were performed under the same solution conditions. Final concentrations were 200 nM ribozyme, 50 mM Na-MOPS, pH 7.0, 5 mM Mg^{2+}, 2 mM Mg^{2+}·ATP (all Mg^{2+} was added as magnesium acetate). From the addition of CYT-19 or equivalent buffer, all reactions also contained 2 mM Tris-HCl pH 8.5, 50 mM KCl, 0.1 mM EDTA, 0.02 mM DTT and 5% (v/v) glycerol.

At various times, aliquots were 'quenched' by threefold dilution into a solution containing a high MgCl$_2$ concentration (50 mM final concentration), which slows the refolding of the misfolded ribozyme to unmeasurable levels[22] and blocks CYT-19 activity (see Supplementary Fig. 1). This solution also included 50 mM Na-MOPS, pH 7.0, 500 µM guanosine, and for reactions of the E^{ΔP5abc} ribozyme it included 250 nM P5abc RNA.

After a further incubation for 5–30 min at 25 °C, the fraction of native ribozyme was determined by adding a trace amount of ^{32}P-labelled substrate (CCCUCUA$_5$). Because the substrate binds both the misfolded and native species of the ribozyme but is cleaved only by the native ribozyme, a burst of product formation is observed; the amplitude provides a measure of the fraction of native ribozyme. For each ribozyme variant, in initial experiments aliquots of this cleavage reaction were quenched at various times thereafter by adding 2 volumes of a solution containing 70% formamide and 100 mM EDTA, thus giving a complete reaction time course. For both the P5a variant and the E^{ΔP5abc} ribozyme on the addition of P5abc[34], cleavage under these conditions was rapid (more than 10 min^{-1}), similar to that of the wild-type ribozyme, and thus the amplitude of the burst was obtained in subsequent experiments by using a single reaction time of 1 min (ref. 25). This time is sufficient for complete cleavage by the native ribozyme but not for significant dissociation of substrate from the misfolded ribozyme[22]. For reactions of the E^{ΔP5abc} ribozyme in complex with CYT-18 protein, cleavage was somewhat slower (about 1–2 min^{-1}; data not shown), and therefore a reaction time of 2 min was used. For all reactions, quenched time points were processed by separating labelled substrate and the shorter oligonucleotide product by 20% denaturing PAGE and quantifying the reaction progress with a phosphorimager.

Measurement of folding parameters for P5a mutant in the absence of CYT-19. For use in simulations of CYT-19-mediated unfolding and subsequent refolding of the P5a mutant, the initial folding parameters were measured by using ribozyme activity as described previously[20]. Tertiary folding (660 nM ribozyme) was initiated by adding Mg^{2+} in the presence of 750 µM guanosine and a small excess of ^{32}P-labelled substrate (1 µM). At various times, reactions were quenched by adding two volumes of a solution containing 70% formamide and 100 mM EDTA, and labelled substrate and product were separated by 20% denaturing PAGE and analysed as above. The progress of the cleavage reaction reflects the progress of folding, because control experiments (not shown) indicated that substrate cleavage by this ribozyme variant is at least 10 min^{-1} at a Mg^{2+} concentration of 10 mM or more, substantially faster than folding under the same conditions. Thus, folding is rate-limiting for cleavage, so the progress of cleavage gives the rate constant for folding and the fraction of the ribozyme that folds directly to the native state, avoiding misfolding (see Scheme 1).

48. Herschlag, D. & Cech, T. R. Catalysis of RNA cleavage by the *Tetrahymena thermophila* ribozyme. 1. Kinetic description of the reaction of an RNA substrate complementary to the active site. *Biochemistry* **29**, 10159–10171 (1990).

LETTERS

A belt of moonlets in Saturn's A ring

Miodrag Sremčević[1], Jürgen Schmidt[2], Heikki Salo[3], Martin Seiß[2], Frank Spahn[2] & Nicole Albers[1]

The origin and evolution of planetary rings is one of the prominent unsolved problems of planetary sciences, with direct implications for planet-forming processes in pre-planetary disks[1]. The recent detection of four propeller-shaped features in Saturn's A ring[2] proved the presence of large boulder-sized moonlets in the rings[3–5]. Their existence favours ring creation in a catastrophic disruption of an icy satellite rather than a co-genetic origin with Saturn, because bodies of this size are unlikely to have accreted inside the rings. Here we report the detection of eight new propeller features in an image sequence that covers the complete A ring, indicating embedded moonlets with radii between 30 m and 70 m. We show that the moonlets found are concentrated in a narrow 3,000-km-wide annulus 130,000 km from Saturn. Compared to the main population of ring particles[6–8] (radius $s < 10$ m), such embedded moonlets have a short lifetime[9] with respect to meteoroid impacts. Therefore, they are probably the remnants of a shattered ring-moon of Pan size or larger[2], locally contributing new material to the older ring. This supports the theory of catastrophic ring creation in a collisional cascade[9–12].

On 20 August 2005 the Cassini spacecraft recorded an occultation of the star α Scorpii (α Sco, Antares) by Saturn's rings. The Imaging Science Subsystem (ISS)[13] Narrow Angle Camera (NAC) took 26 images with a complete radial coverage of Saturn's A ring. Four of these images show seven propeller features, all radially inward from the Encke division, which is also where the first propellers were detected[2] in images obtained during the Saturn Orbit Insertion (SOI) manoeuvre of Cassini. Shortly after the α Sco sequence, NAC recorded 105 images of the Encke gap. In this 'movie' sequence, one additional propeller is found 340 km outside the Encke division. Altogether we can identify eight new moonlet-induced structures (see Supplementary Information for details; Figs 1, 2

and Supplementary Figs 1–9); of these eight, five re-occur in subsequent images, confirming their orbital motion (Fig. 1).

All propellers were identified in images of the backlit rings, so that dark features could correspond either to very dense regions (totally blocking the light) or to mostly empty regions (not scattering any light). Similarly, the bright propeller streaks could be either density depletions or density enhancements. Our photometric modelling of simulated embedded moonlets and ring background (Supplementary Figs 12–14) strongly favours the correspondence of bright parts of the propeller with increased density. Taking into account the unresolved self-gravity wakes in ring A[14–16], which reduce the brightness compared to a uniform ring[2,17], we can account qualitatively for the ring background I/F (where I is brightness and F is the solar flux) for α Sco and SOI image geometries (Supplementary Fig. 12). However, it is still a challenge to explain quantitatively the large brightness difference between the propellers and the background. A potential factor contributing to propeller brightness could be the release of debris from regolith-covered ring particles[18,19] in the moonlet-perturbed region, where the relative impact speeds of particles are considerably larger than in the background ring. Our models (Supplementary Figs 13 and 14) imply that this debris can significantly enhance the brightness of the moonlet-induced wakes—the waves induced in the ring downstream from the moonlet by its gravity[20,21]—which we suggest are seen in images as propellers.

We can determine the propeller dimensions by fitting the brightness I/F of the wings to a double-gaussian function (Fig. 2 and Supplementary Information). The observed (Fig. 3) azimuthal (W) and radial (h) sizes rule out the cubic scaling $W \propto h^3$ predicted theoretically[4,5,22] for the length of an incomplete gap opened by a moonlet. Instead, a nearly linear relationship $W \propto h^\beta$, where $\beta \approx 1-1.5$, is indicated.
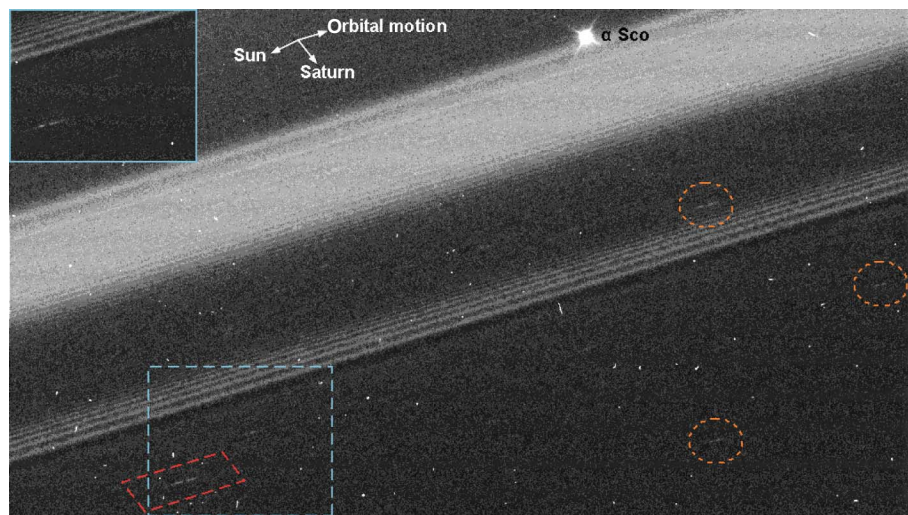

Figure 1 | Four new propellers in Saturn's A ring seen by Cassini. Part of the image N1503229987 (Supplementary Figs 1 and 2) with four propellers (enclosed in different colours to facilitate comparison). The greyscale colour represents I/F in the range of 0.0071–0.0167. The inset demonstrates the re-occurrence of the largest propeller (blue dashed rectangle) in the image N1503230047 (orbital motion is subtracted; Supplementary Figs 1 and 3). The images are 2 of 26 in the α Sco sequence with a complete radial coverage of the A ring and part of the Cassini division. In this sequence, NAC observed the unlit side of the rings, with a phase angle of 126° and incidence and emission angles of 111° and 58°, respectively. The images were taken at 60 s intervals with 50 ms exposure in a clear filter with a resolution of 1 km per pixel in the radial direction and 0.5 km per pixel in the azimuthal direction that is only superseded by the SOI image sequence[2,13].

[1]Laboratory for Atmospheric and Space Physics, University of Colorado at Boulder, 392 UCB, Boulder, Colorado 80309-0392, USA. [2]Nonlinear Dynamics, Department of Physics, University of Potsdam, Am Neuen Palais 10, 14469 Potsdam, Germany. [3]Astronomy Division, Department of Physical Sciences, University of Oulu, 90014 Oulu, Finland.

A nearly linear scaling is expected for the size of the wake region, because wakes are damped (independently of the moonlet size) after a certain number of wake crests[20,22,23] (see Supplementary Information). This supports our interpretation that the bright streaks seen in images are actually related to moonlet wakes. A deviation from the linear trend follows naturally if the propeller brightness is further enhanced by release of regolith debris in the wake region (Fig. 3).

All propellers in the α Sco and SOI images are found in an annulus between 128,700 and 131,700 km from Saturn. A homogeneous distribution of moonlets over the whole A ring is extremely unlikely with the given observation (a probability of 10^{-6} from poissonian statistics applied to the α Sco image sequence). One propeller-moon orbits at a saturnian distance of 134,000 km outside the Encke division (Fig. 2b). Thus, moonlets are found preferentially inward from the Encke division, and are significantly less frequent elsewhere in the A ring. The moonlet sizes are estimated from the propeller dimensions, and range from 30 to 70 m in radius (Fig. 4). Identifying the brightness enhancements in images as moonlet wakes, we re-interpret the SOI sequence moonlet radii to be about 20 m.

The resulting cumulative moonlet size distribution (Fig. 4) shows a steep slope of index $Q = 10$, the fairly large error bars requiring a minimal slope of $Q > 8$. The non-discovery of propellers in the α Sco images of the rest of the A ring implies a very low frequency there or locally, even the total absence of objects larger than 10 m. Moreover, from the size distribution we expect no moonlets larger than about 350 to 500 m in the moonlet region, which is consistent with the non-detection of circumferential gaps[5] apart from the Keeler and Encke gaps. It is also possible that Pan and Daphnis, together with putative non-detected larger moonlets in the Encke gap[13], form a shallower wing of the size distribution[24].

The inhomogeneous distribution of moonlets in the A ring and their rather steep size distribution have interesting consequences for the formation and evolution of Saturn's rings. It has been suggested

that the embedded moonlets favour a scenario of ring creation in a break-up of a larger body[2]. However, it seems unlikely that moonlets are remainders of a single catastrophic event that created the whole ring system, because in this case a uniform distribution would emerge. Instead, the moonlet belt is compatible with a more recent break-up of a body orbiting in the A ring. Accretion of moonlets from a population of smaller ring particles seems equally unlikely: even if the absence of accretion radially inward from a radial distance of 128,000 km from Saturn could be attributed to the increasing tidal forces nearer to the planet[25], it would be difficult to understand why no moonlets accreted in the outer A ring. The probability of finding coincidentally no propellers outside 132,000 km in the α Sco images, if moonlets were uniformly distributed, is still less than 5×10^{-4}. Furthermore, the fact that one propeller resides in the Prometheus 12:11 density-wave-train and another in the Pan-induced wakes (Fig. 1 and Fig. 2b) demonstrates that propellers are not necessarily destroyed in perturbed ring regions. Hence, the possibility that the outer A ring is too perturbed to harbour such moonlets also seems unlikely.

Catastrophic collisions are a common concept in the theory of ring evolution, and shattering of larger moons in a collisional cascade has been proposed[9–12]. In this scenario the disruption of larger fragments at a later time gradually adds fresh material to the existing ring system. Evidence for a very recent (1984) disrupting impact of a metre-sized object on an icy boulder of similar size in Saturn's D ring is given by the detection of a rapidly evolving tightly wound spiral pattern[26]. Moreover, a number of potential target moons of kilometre size still exist at present in the rings (Pan, Daphnis, Atlas), and smaller moons are expected in the gaps[13]. Consequently, we may expect to

Figure 2 | Close-up view and re-projection of two propeller structures.
a, The red rectangle from Fig. 1; **b**, the portion of the image N1503243458 (image in Encke gap 'movie' sequence covering 44° of the circumference with resolution better than 2 km per pixel; first-order Pan wakes are visible in the background, Supplementary Fig. 6). Structural fits are indicated: $2W = 2w + 2a$ is the total azimuthal extent and $2h$ is the radial separation of the propeller wings (corresponds to Δr in ref. 2), whereas (a, b) denote the semi-major axes of ellipses fitted to the wings (dots). In **a**, $a_0 = 131,525$ km; in **b**, $a_0 = 134,079$ km. Orbital motion is to the right.
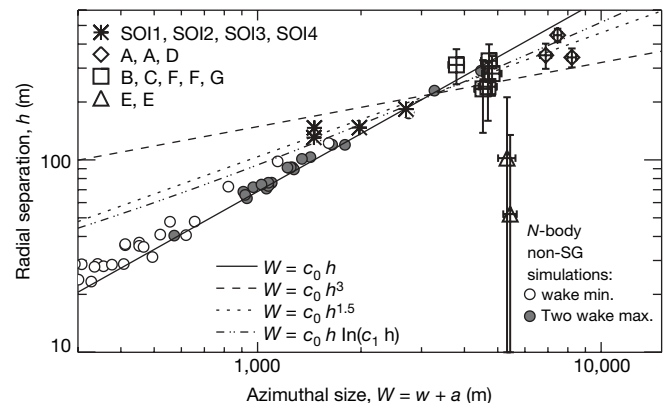
Figure 3 | Nearly linear spatial scaling of the features points to moonlet wakes. Shown is the radial separation h of the propeller wings as a function of their azimuthal extent $W = w + a$. Asterisks represent propellers in SOI images[2]; symbols with $\pm 1\sigma$ error bars (A–E, Supplementary Table 2) denote seven new propellers from the α Sco sequence. Various fits to the data are indicated, excluding the anomalous points E; the cubic relation $W \propto h^\beta$ where $\beta = 3$ is clearly ruled out by the data, which seem to suggest $\beta \approx 1 – 1.5$. This nearly linear scaling implies that propellers are related to the moonlet wakes, whose azimuthal wavelength $\lambda_1 = 3\pi h$ increases linearly with radial distance $h = |r - a_0|$ to the moonlet. A linear scaling $W \propto \lambda_1 \propto h$ would be obtained if the wakes are damped after a certain number of wake-crests, independently of the moon size. Such an enhanced localized damping is indeed expected near the point of streamline crossing[20], roughly two wake cycles downstream from the moon, which is confirmed by N-body simulations[22,23] (Supplementary Fig. 10; here we plot results from non-self-gravitating (non-SG) simulations). The excess to the linear scaling is also in accordance with the picture that the visibility of propellers is partly caused by release of debris in high-speed impacts in the region of moonlet-induced wakes. Namely, with increased moonlet size, enhanced debris production per surface area is expected. Let us assume an exponential downstream decay of the optical depth of debris caused by re-accumulation. Then, a critical minimum optical depth of debris is required for the propeller to stand out against the background, leading to a logarithmic dependence between azimuthal extent and debris release rate. Furthermore, assuming that debris production scales as some power of moonlet mass suggests $W \propto h\ln(c_1 h)$.
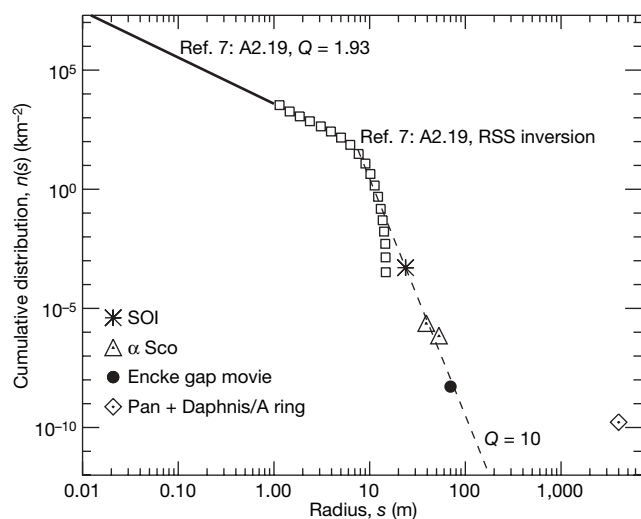
Figure 4 | Cumulative size distribution of particles in the moonlet belt region. The Voyager Radio Science Subsystem (RSS) results[7] have a knee around $s = 10$ m, indicating a steeper slope for $s > 10$ m, the latter points having an order of magnitude uncertainty (A2.19 refers to the A-ring region studied by RSS). Individual moonlet sizes are derived from the spatial scaling of the propellers. The largest density enhancement in a moonlet wake lies at a radial distance[5,22,30] $h/H \approx 4.5 \pm 0.5$, where $H = a_0[M/(3M_{Saturn})]^{1/3}$ is the moonlet's Hill scale, which is proportional to its radius. M denotes moonlet mass and M_{Saturn} denotes Saturn's mass. Assuming densities between 0.5 and 0.9 g cm^{-3}, we may estimate the moonlet radius as $s = 0.14h$ with a probable error of about 15% from the radial separation of the propeller wings. Assuming the whole A ring as basis for calculation (instead of only the moonlet belt within $r_1 = 128,500$ km and $r_2 = 134,500$ km) would decrease α Sco values by a factor of three, and, likewise, a narrower moonlet belt would increase α Sco values by a factor of two at most. Other points are affected even less. Additionally, poissonian statistics of propeller appearances in the images implies a statistical error of 50% for SOI, <50% for α Sco and 100% for the Encke movie sequence. The latter point has additional uncertainty owing to long image exposure (2 s) and possible smear, which might hide propellers of similar size. However, all these factors are small and do not change the overall picture nor the inferred steepness of the distribution. Note that identifying bright features as gaps would just imply systematically two-times larger sizes s, not affecting the overall conclusions.

find traces of the past break-up of larger objects in the current rings[12]. Steep size-slopes between $Q = 5$ and $Q = 8$ have been reported for the fragments of giant impacts[24,27]. Combining all moonlets ($s > 15$ m) in the belt, we obtain a sphere of roughly 10 km in radius, suggesting that a moon of Pan size or larger has been shattered. With a mean ejection speed of 50 ms^{-1} the debris is spread over a radial width of 3,000 km (see Supplementary Information). Using a ring viscosity[28] of $v = 90$ cm^2 s^{-1} and a belt width $\Delta r = 3,000$ km, we obtain a characteristic time $t_{visc} = (\Delta r)^2/v$ of 3×10^7 years, which should give an upper limit for closing the gap previously kept open by the moon. The debris evolves by further shattering resulting from meteoroid bombardment. Whereas larger fragments are gradually ground down to smaller sizes, their size-slope steepening with time, the distribution of ring particles smaller than about 10 m is stabilized by a balance between aggregation and disintegration[18,19]. We estimate the time for the destruction of all moonlets (see Supplementary Information) larger than 100 m in the belt to be 10^8 years (threefold for moonlets >50 m), although the uncertainties of the model[9,29] imply large errors. Thus, the inferred steepness of the moonlets' size distribution and their apparent depletion in the rest of the A ring represent different phases of the process of moonlet destruction, like a clock displaying the age of a ring region.

Note added in proof: Another study of A-ring propellers using a larger set of Cassini data has been recently submitted to *The Astronomical Journal*[31].

1. Burns, J. A. & Cuzzi, J. N. Our local astrophysical laboratory. *Science* **312**, 1753–1755 (2006).
2. Tiscareno, M. S. *et al.* 100-metre-diameter moonlets in Saturn's A ring from observations of propeller structures. *Nature* **440**, 648–650 (2006).
3. Julian, W. H. & Toomre, A. Non-axisymmetric responses of differentially rotating disks of stars. *Astrophys. J.* **146**, 810–830 (1966).
4. Spahn, F. & Sremčević, M. Density patterns induced by small moonlets in Saturn's rings? *Astron. Astrophys.* **358**, 368–372 (2000).
5. Sremčević, M., Spahn, F. & Duschl, W. J. Density structures in perturbed thin cold discs. *Mon. Not. R. Astron. Soc.* **337**, 1139–1152 (2002).
6. Marouf, E. A., Tyler, G. L., Zebker, H. A., Simpson, R. A. & Eshleman, V. R. Particle size distributions in Saturn's rings from Voyager 1 radio occultation. *Icarus* **54**, 189–211 (1983).
7. Zebker, H. A., Marouf, E. A. & Tyler, G. L. Saturn's rings — particle size distributions for thin layer model. *Icarus* **64**, 531–548 (1985).
8. Nicholson, P. *et al.* Saturns rings I: optical depth profiles from the 28 sgr occultation. *Icarus* **145**, 474–501 (2000).
9. Colwell, J. E., Esposito, L. W. & Bundy, D. Fragmentation rates of small satellites in the outer solar system. *J. Geophys. Res.* **105**, 17589–17600 (2000).
10. Esposito, L. W. & Colwell, J. E. Creation of the Uranus rings and dust bands. *Nature* **339**, 605–607 (1989).
11. Colwell, J. E. & Esposito, L. W. Origins of the rings of Uranus and Neptune. I. Statistics of satellite disruptions. *J. Geophys. Res.* **97**, 10227–10241 (1992).
12. Esposito, L. W. *et al.* Ultraviolet imaging spectroscopy shows an active saturnian system. *Science* **307**, 1251–1255 (2005).
13. Porco, C. C. *et al.* Cassini imaging science: initial results on Saturn's rings and small satellites. *Science* **307**, 1226–1236 (2005).
14. Dones, L., Cuzzi, J. N. & Showalter, M. R. Voyager photometry of Saturn's A ring. *Icarus* **105**, 184–215 (1993).
15. Colwell, J. E., Esposito, L. W. & Sremčević, M. Self-gravity wakes in Saturn's A ring measured by stellar occultations from Cassini. *Geophys. Res. Lett.* **33**, 7201–7204 (2006).
16. Hedman, M. M. *et al.* Self-gravity wake structures in Saturn's A ring revealed by Cassini VIMS. *Astron. J.* **133**, 2624–2629 (2007).
17. Salo, H., Karjalainen, R. & French, R. G. Photometric modeling of Saturn's rings. II. Azimuthal asymmetry in reflected and transmitted light. *Icarus* **170**, 70–90 (2004).
18. Weidenschilling, S. J., Chapman, C. R., Davis, D. R. & Greenberg, R. in *Planetary Rings* (eds Greenberg, R. & Brahic, A.) 367–415 (Univ. Arizona Press, Tucson, 1984).
19. Albers, N. & Spahn, F. The influence of particle adhesion on the stability of agglomerates in Saturn's rings. *Icarus* **181**, 292–301 (2006).
20. Showalter, M. R., Cuzzi, J. N., Marouf, E. A. & Esposito, L. W. Satellite 'wakes' and the orbit of the Encke Gap moonlet. *Icarus* **66**, 297–323 (1986).
21. Showalter, M. R. Visual detection of 1981S13, Saturn's eighteenth satellite, and its role in the Encke gap. *Nature* **351**, 709–713 (1991).
22. Seiß, M., Spahn, F., Sremčević, M. & Salo, H. Structures induced by small moonlets in Saturn's rings: implications for the Cassini Mission. *Geophys. Res. Lett.* **32**, 11205–11208 (2005).
23. Lewis, M. C. & Stewart, G. R. Features around embedded moonlets in Saturn's rings: the role of self-gravity and particle size distributions. *Icarus* (submitted).
24. Durda, D. D. *et al.* Size frequency distributions of fragments from SPH/N-body simulations of asteroid impacts: comparison with observed asteroid families. *Icarus* **186**, 498–516 (2007).
25. Karjalainen, R. & Salo, H. Gravitational accretion of particles in Saturn's rings. *Icarus* **172**, 328–348 (2004).
26. Hedman, M. M. *et al.* Saturn's dynamic D ring. *Icarus* **188**, 89–107 (2007).
27. Michel, P., Benz, W. & Richardson, D. C. Disruption of fragmented parent bodies as the origin of asteroid families. *Nature* **421**, 608–611 (2003).
28. Daisaka, H., Tanaka, H. & Ida, S. Viscosity in a dense planetary ring with self-gravitating particles. *Icarus* **154**, 296–312 (2001).
29. Cuzzi, J. N. & Estrada, P. R. Compositional evolution of Saturn's rings due to meteoroid bombardment. *Icarus* **132**, 1–35 (1998).
30. Spahn, F. & Sponholz, H. Existence of moonlets in Saturn's rings inferred from the optical depth profile. *Nature* **339**, 607–608 (1989).
31. Tiscareno, M. S., Burns, J. A., Hedman, M. M. & Porco, C. C. The population of propellers in Saturn's A ring. *Astron. J.* (submitted).

LETTERS

Discovery of ^{40}Mg and ^{42}Al suggests neutron drip-line slant towards heavier isotopes

T. Baumann[1], A. M. Amthor[1,2], D. Bazin[1], B. A. Brown[1,2], C. M. Folden III[1], A. Gade[1,2], T. N. Ginter[1], M. Hausmann[1], M. Matoš[1], D. J. Morrissey[1,3], M. Portillo[1], A. Schiller[1], B. M. Sherrill[1,2], A. Stolz[1], O. B. Tarasov[1,4] & M. Thoennessen[1,2]

A fundamental question in nuclear physics is what combinations of neutrons and protons can make up a nucleus. Many hundreds of exotic neutron-rich isotopes have never been observed; the limit of how many neutrons a given number of protons can bind is unknown for all but the lightest elements[1], owing to the delicate interplay between single particle and collective quantum effects in the nucleus. This limit, known as the neutron drip line, provides a benchmark for models of the atomic nucleus. Here we report a significant advance in the determination of this limit: the discovery of two new neutron-rich isotopes—^{40}Mg and ^{42}Al—that are predicted to be drip-line nuclei[2]. In the past, several attempts to observe ^{40}Mg were unsuccessful[3,4]; moreover, the observation of ^{42}Al provides an experimental indication that the neutron drip line may be located further towards heavier isotopes in this mass region than is currently believed. In stable nuclei, attractive pairing forces enhance the stability of isotopes with even numbers of protons and neutrons. In contrast, the present work shows that nuclei at the drip line gain stability from an unpaired proton, which narrows the shell gaps and provides the opportunity to bind many more neutrons[5,6].

The experimental determination of the neutron drip line is extremely challenging primarily because isotopes at the drip line, for example, perhaps 18 neutrons beyond the stable aluminium isotope, can be produced only in minute quantities if they are accessible at all. The other difficulty is that these nuclei are very fragile, and they

rapidly and preferentially emit neutrons when produced in any nuclear reaction. The theoretical prediction of the drip line is also very difficult because of the lack of a reliable universal theory of the atomic nucleus. From the many theoretical models that are available we selected two of the best global models, the finite range droplet model (FRDM) and the Hartree–Fock–Bogoliubov model. Figure 1 shows the predicted neutron drip line from these two models and illustrates the large variation in model predictions. The FRDM uses a semiclassical description of the macroscopic contributions to the nuclear binding energy augmented with microscopic corrections resulting from local single-particle shell structure and the pairing of nucleons[7] (Fig. 1, solid black line). For comparison, the fully microscopic Hartree–Fock–Bogoliubov model (HFB-8) is a state-of-the-art quantum mechanical calculation[8] that puts the nucleons into a mean field with a Skyrme interaction in which the pairing field is included in analogy to the Bogoliubov–de Gennes equations in condensed matter physics[9] (Fig. 1, dashed green line). We selected HFB-8 over other HFB models because it has the best overall fit to measured masses[10]. Although both FRDM and HFB-8 correctly predict the location of the neutron drip line in many cases, they are not able to account for the detailed interplay of valence protons and neutrons. The discrepancies between the models are particularly large in the region from magnesium to silicon. This issue is especially relevant to the current effort to determine the limits of nuclear existence and the ability to determine those limits with the next generation of nuclear
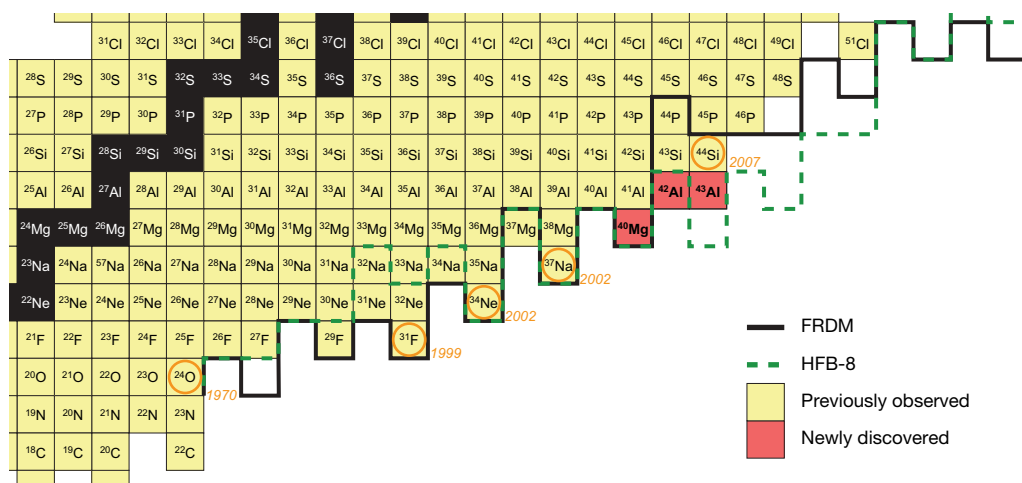

Figure 1 | Section of the chart of nuclides for light, neutron-rich nuclei. The proton number increases vertically and the neutron number horizontally. Yellow squares denote previously observed nuclei. The neutron drip lines predicted by the FRDM and HFB-8 models are shown by the black and

dashed green lines, respectively. The most recently observed drip-line nuclei are indicated by orange circles with their year of discovery, and the isotopes discovered in the present experiment are highlighted in red (see the text for details).

[1]National Superconducting Cyclotron Laboratory, [2]Department of Physics and Astronomy, [3]Department of Chemistry, Michigan State University, East Lansing, Michigan 48824, USA. [4]Flerov Laboratory of Nuclear Reactions, JINR, 141980 Dubna, Moscow region, Russian Federation.

research facilities such as the Facility for Antiproton and Ion Research (FAIR) in Europe, the Radioisotope Beam Factory (RIBF) in Japan and the Facility for Rare-Isotope Beams (FRIB) in the United States.

The current knowledge of the drip line is limited to only the lightest nuclei. The last bound oxygen isotope, ^{24}O (atomic number $Z = 8$), was observed in 1970 (ref. 11). However, the neutron drip line was established for oxygen only in 1997 after the nuclei ^{25}O to ^{28}O had been shown to be unbound with respect to prompt neutron emission[12–14]. Subsequently, the isotopes ^{31}F, ^{34}Ne and ^{37}Na were observed[3,4], and although it is generally believed that they lie along the drip line, the fact that ^{33}F, ^{36}Ne and ^{39}Na are indeed unbound has not been experimentally established. These experiments also failed to observe the even–even nucleus ^{40}Mg and it was speculated that ^{40}Mg might be unbound[15]. Here we report the production and identification of ^{40}Mg for the first time, as well as that of the neighbouring odd-Z nucleus ^{42}Al.

The fragmentation of stable nuclei followed by the separation and identification of the products in less than 1 μs is the only method currently available to produce nuclei at or near the neutron drip line[1]. The observation of the isotopes of interest requires a high primary-beam intensity, a high collection efficiency, a high efficiency of identification and—perhaps most importantly—a very high degree of purification because they are produced very rarely, in only about 1 in 10^{15} reactions. These requirements are met at the National Superconducting Cyclotron Laboratory[16] at Michigan State University with the use of the A1900 fragment separator[17] and S800 analysis beamline[18]. A primary beam of ^{48}Ca with an energy per nucleon of 141 MeV was interacted with a ^{nat}W target, 970 μm thick, placed at the object position of the A1900 fragment separator (see Fig. 2). The near beam-velocity products along with the unreacted primary beam particles were separated first by their momentum-to-charge ratio in the A1900, which accepted a ±2.5% momentum deviation and had an angular acceptance of 120 mrad by 80 mrad, resulting in a total acceptance of about 23% of the yield. The momentum-selected fragments then passed through an aluminium degrader, 156 μm thick, at the momentum-dispersive midplane that preserved the achromaticity of the separator while providing a physical separation of the remaining reaction products by nuclear charge. The momentum and the isotopic identity of the selected particles (about 1 in 5×10^{10}) were determined in the S800 analysis system on an event-by-event basis by position measurement of the momentum-dispersed ions. A detector stack consisting of a position-sensitive parallel-plate avalanche counter, seven silicon PIN diodes, 1 mm thick, and a plastic scintillator, 50 mm thick, provided the particle identification, which required consistent energy-loss measurements in all seven silicon detectors and consistent determination of two

flight times over 21 m and 46 m flight paths. After initial calibrations the entire system was set to the expected magnetic rigidity for ^{40}Mg, corresponding to 4.78 T m after the production target. Data were collected for a total of 7.6 days over an 11-day period; the average incident beam intensity was 5.0×10^{11} particles per second.

The particle identification can be seen in Fig. 3, where the locus of isotopes with constant $N = 2Z$ is indicated by the vertical line and heavier isotopes lie to the right. Three events of ^{40}Mg were clearly identified. Each of the parameters that are used for the particle identification has been checked on an event-by-event basis to exclude possible ambiguous background events. The non-observation of any events that would correspond to ^{39}Mg indicates that it is unbound. The probability that the three events identified as ^{40}Mg correspond to ^{38}Mg, on the basis of the distribution of ^{38}Mg events, is less than 3×10^{-15}. In addition, many events of the previously observed ^{31}F, ^{34}Ne and ^{37}Na isotopes were detected. The fact that ^{30}F, ^{33}Ne and ^{36}Na are unbound (lack of events) was also confirmed. Further, the 23 events of ^{42}Al establish its discovery. Figure 3 also contains one event consistent with ^{43}Al. We determined a probability of 2.4×10^{-3} that this event was caused by a possible contribution from the neighbouring ^{42}Al. Furthermore, as a result of the attractive neutron-pairing interaction, the firm observation of the odd–odd isotope ^{42}Al$_{29}$ supports the existence of ^{43}Al$_{30}$, lending credibility to the interpretation of the single event as evidence for the existence of this nucleus. Our experiment also confirmed the recent observation of ^{44}Si with the A1900 separator[19].

The discovery of the even–even ^{40}Mg isotope is consistent with the predictions of both the FRDM and the HFB-8 model and with the staggered pattern of the drip line in this region. It is interesting to note that if the even–even ^{40}Mg$_{28}$ had not been observed in the present experiment, the drip line might have been considered determined up to magnesium. However, with the observation of ^{40}Mg the question remains open as to whether ^{31}F, ^{34}Ne, ^{37}Na and ^{40}Mg are the last bound isotopes of fluorine, neon, sodium and magnesium, respectively.

More important than the observation of the even–even ^{40}Mg is the discovery of the odd–odd ^{42}Al, which was predicted to be unbound by the FRDM and HFB-8 models. The present observation breaks the pattern of the staggering at the drip line. The addition of one unpaired proton in the $0d_{5/2}$ orbital to magnesium seems to have a significant influence on the stability of the most neutron-rich isotopes, similar to the transition from oxygen to fluorine. Quantum-mechanical calculations such as HFB-8 and the $sdfp$-shell model[20] predict that the $1p_{3/2}$ orbital will start to fill at $N = 29$ and that the single-particle energy of the $1p_{3/2}$ orbital remains more or less constant as a function of neutron number. Our observation of ^{42}Al therefore suggests that all of the aluminium isotopes up to the
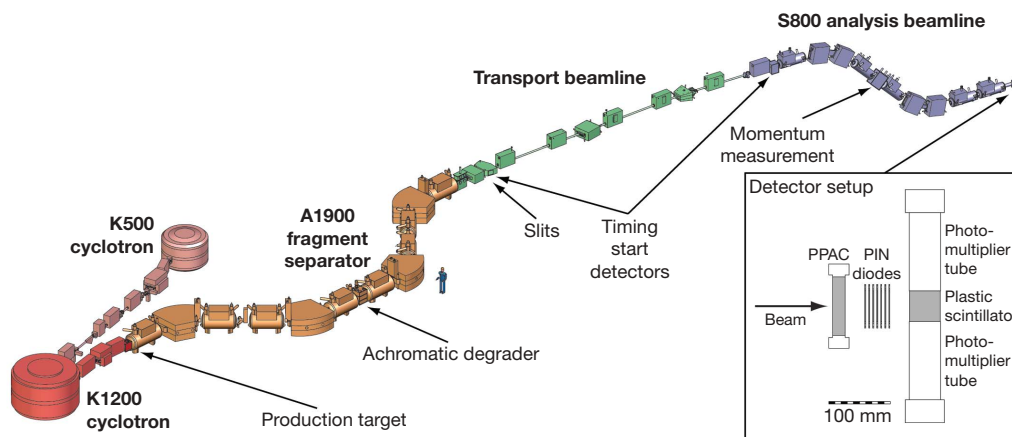

Figure 2 | Schematic diagram of the coupled cyclotron facility[16]. This overview diagram shows the major components used for the production,

separation and identification of new neutron-rich nuclei. PPAC, parallel-plate avalanche counter.
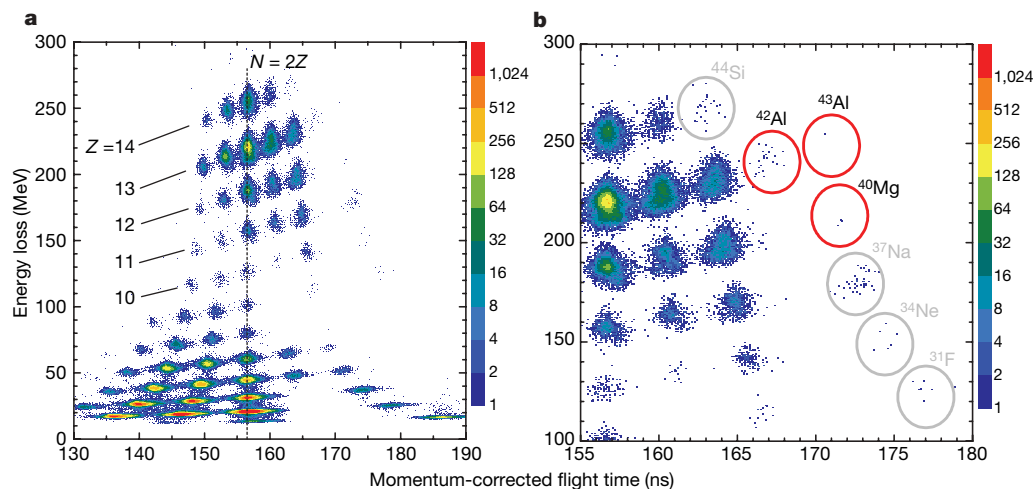
Figure 3 | Particle identification plots of observed isotopes. a, Particle identification of the neutron-rich products between beryllium and silicon, including the new isotopes ^{40}Mg, ^{42}Al and the single event identified as ^{43}Al. **b,** Expanded region of heaviest exotic nuclei. The energy loss in the first PIN detector is plotted against the flight time of the nuclei through the S800

analysis line, corrected to the momentum of the central path, indicating the unambiguous identification of each isotope. Isotopes of an element lie on diagonal lines and isotopes with $N = 2Z$ lie along a single vertical line (dashed line in **a**).

nominal filling of the $1p_{3/2}$ orbital (^{45}Al$_{32}$) are most probably bound. Heavier nuclei up to ^{47}Al are also probably bound if the $1p_{3/2}$–$1p_{1/2}$ spin–orbital splitting becomes small, as suggested by the new generation of Skyrme forces that include the tensor interaction[5,6]. This is illustrated by a more recent version of the HFB calculation (HFB-9), which does predict ^{42}Al to be bound and actually shows even ^{48}Al to be bound[10]. This demonstrates the likelihood that the drip line is significantly further from stability than predicted by models such as HFB-8, which predict that ^{42}Al is unbound.

Thus, the observation of ^{40}Mg, ^{42}Al and possibly even ^{43}Al at or near the neutron drip line provides the first experimental indication that the stability of very neutron-rich aluminium isotopes is enhanced relative to magnesium. This provides a significant advance in our understanding of where the drip line is likely to lie in this region and indicates that it may be farther from stable isotopes than expected. Hence, the goal of defining the limits of stability at the current and next generations of nuclear science facilities built to study rare isotopes may be significantly more difficult than previously believed. In fact, the results of this experiment indicate that the drip line may be beyond the reach of all of these facilities for nuclei with atomic numbers greater than 12.

1. Thoennessen, M. Reaching the limits of nuclear stability. *Rep. Prog. Phys.* **67,** 1187–1232 (2004).
2. Audi, G., Wapstra, A. H. & Thibault, C. The AME2003 atomic mass evaluation. *Nucl. Phys. A* **729,** 337–676 (2003).
3. Notani, M. *et al.* New neutron-rich isotopes, ^{34}Ne, ^{37}Na and ^{43}Si, produced by fragmentation of a 64 A MeV ^{48}Ca beam. *Phys. Lett. B* **542,** 49–54 (2002).
4. Lukyanov, S. M. *et al.* Experimental evidence for the particle stability of ^{34}Ne and ^{37}Na. *J. Phys. G Nucl. Part. Phys.* **28,** L41–L45 (2002).
5. Otsuka, T., Matsuo, T. & Abe, D. Mean field with tensor force and shell structure of exotic nuclei. *Phys. Rev. Lett.* **97,** 162501 (2006).
6. Brown, B. A., Duguet, T., Otsuka, T., Abe, D. & Suzuki, T. Tensor interaction contributions to single-particle energies. *Phys. Rev. C* **74,** 061303(R) (2006).
7. Möller, P., Nix, J. R., Myers, W. D. & Swiatecki, W. J. Nuclear Ground-State Masses and Deformations. *At. Data Nucl. Data Tables* **59,** 185–381 (1995).
8. Samyn, M., Goriely, S., Bender, M. & Pearson, J. M. Further explorations of Skyrme-Hartree-Fock-Bogoliubov mass formulas. III. Role of particle-number projection. *Phys. Rev. C* **70,** 044309 (2004).
9. de Gennes, P. G. *Superconductivity of Metals and Alloys* (Addison-Wesley, Reading, Massachusetts, 1989).
10. Goriely, S., Samyn, M., Pearson, J. M. & Onsi, M. Further explorations of Skyrme–Hartree–Fock–Bogoliubov mass formulas. IV: Neutron-matter constraint. *Nucl. Phys. A* **750,** 425–443 (2005).
11. Artukh, A. G. *et al.* New isotopes ^{21}N, ^{23}O, ^{24}O and ^{25}F, produced in nuclear reactions with heavy ions. *Phys. Lett.* **32B,** 43–44 (1970).
12. Langevin, M. *et al.* Production of neutron-rich nuclei at the limits of particles stability by fragmentation of 44 MeV/u ^{40}Ar projectiles. *Phys. Lett.* **150B,** 71–74 (1985).
13. Guillemaud-Mueller, D. *et al.* Particle stability of the isotopes ^{26}O and ^{32}Ne in the reaction 44 MeV/nucleon ^{48}Ca + Ta. *Phys. Rev. C* **41,** 937–941 (1990).
14. Tarasov, O. *et al.* Search for ^{28}O and study of neutron-rich nuclei near the $N = 20$ shell closure. *Phys. Lett. B* **409,** 64–70 (1997).
15. Lutostansky, Yu. S., Lukyanov, S. M., Penionzhkevich, Yu. E. & Zverev, M. V. Neutron drip line in the region of O–Mg isotopes. *Particles Nuclei Lett.* **115,** 86–93 (2002).
16. Gelbke, C. K. Rare isotope research capabilities at the NSCL today and at RIA in the future. *Prog. Particle Nucl. Phys.* **53,** 363–372 (2004).
17. Morrissey, D. J., Sherrill, B. M., Steiner, M., Stolz, A. & Wiedenhoever, I. Commissioning the A1900 projectile fragment separator. *Nucl. Instrum. Meth. B* **204,** 90–96 (2003).
18. Bazin, D., Caggiano, J. A., Sherrill, B. M., Yurkon, J. & Zeller, A. The S800 spectrograph. *Nucl. Instrum. Methods Phys. Res. B* **204,** 629–633 (2003).
19. Tarasov, O. B. *et al.* New isotope ^{44}Si and systematics of the production cross sections of the most neutron-rich nuclei. *Phys. Rev. C* **75,** 064613 (2007).
20. Retamosa, J., Caurier, E., Nowacki, F. & Poves, A. Shell model study of the neutron-rich nuclei around $N = 28$. *Phys. Rev. C* **55,** 1266–1274 (1997).

Probable heat capacity signature of the supersolid transition

X. Lin[1], A. C. Clark[1] & M. H. W. Chan[1]

Liquid ^{4}He enters the superfluid state and flows without friction below 2.176 K. Thin liquid films adsorbed on solid substrates undergo the same transformation, although at a lower temperature. When the substrate is subjected to oscillatory motion a portion of the film, known as the superfluid fraction, decouples from the oscillation. A similar phenomenon has been observed[1,2] in solid ^{4}He, in which a fraction of the solid seems to decouple from the motion of the surrounding lattice. Although this observation has been replicated in various laboratories[3–6], no thermodynamic signature of the possible supersolid transition has been seen. Here we report the finding of a heat capacity peak that coincides with the onset of mass decoupling. This complementary experimental evidence supports the existence of a genuine transition between the normal solid and supersolid phases of ^{4}He.

The superfluid-like behaviour—that is, nonclassical rotational inertia[7] (NCRI)—of solid ^{4}He appears as a decrease in the period of a sample cell under torsional motion. Relative to the total amount of ^{4}He in the torsion cell, the NCRI fraction (NCRIF) ranges[8] from 0.03% to as high as 20%. The variation suggests that crystalline defects within the solid greatly influence the phenomenon[3,9,10]. Most of the samples studied so far were grown with commercial high-purity ^{4}He gas (~0.3 p.p.m. ^{3}He) by the blocked-capillary method[10]. The temperature dependence of NCRIF, characterized by saturation at low temperature and a gradual decay to zero at high temperature, is qualitatively reproduced in all measurements[1–6,8,10]. The onset temperature T_o, the point at which NCRI becomes resolvable from the noise, varies between 160 and 400 mK. In samples diluted with higher concentrations of ^{3}He, T_o is found to extend to even higher temperature, both in confined geometries[1] and in the bulk (E. Kim, J. S. Xia, J. T. West, X.L. and M.H.W.C., unpublished observations). A recent model equates NCRI to the rotational susceptibility of a vortex liquid phase[11]. According to this model, the high-temperature tail of NCRI reflects the finite response time of vortices to the oscillating flow fields generated by the torsional oscillator. The coupling of ^{3}He atoms to the cores of such vortices will further slow their response, because they must be dragged through the lattice. The consequence of this interaction is to extend the apparent onset to higher temperature, as observed. In addition, it was found[10] that for solid ^{4}He containing ~1 p.p.b. ^{3}He the NCRIF data of eight different samples collapse onto a single curve with a distinctly sharper onset temperature, $T_o = 75 \pm 5$ mK (mean ± s.d). Last, a frequency effect[6] consistent with this theoretical model has also been observed.

If the observed NCRI is the signature of a genuine second-order phase transition, there should be a peak in heat capacity at the transition. Many measurements[12–17] of the heat capacity of solid ^{4}He have been made since the 1960s, all reporting the T^3 dependence expected for a Debye solid at low temperature. However, the resolution of these experiments below 100 mK is severely limited by the heat

capacity of the sample cells, which is typically more than 20-fold that of solid ^{4}He. (See Supplementary Discussion for further experimental details.)

Our sample cell consists of undoped silicon, a small aluminium cap and a minimal amount of epoxy (see inset to Fig. 1). The virtue of silicon is its small heat capacity and high thermal conductivity at low temperature. The aluminium cap is necessary to create a reliable seal between the capillary and silicon. The cell is filled through a thin glass capillary with an inner diameter of 0.1 mm, and is mechanically secured to the refrigerator with nylon screws. Measurements were made with an alternating-current method[18], where the frequency of the applied power is low enough to maintain thermal equilibrium between the entire sample cell and solid helium, but high enough for the alternating-current modulation not to be 'lost' through the weak thermal link (a copper wire 5 cm in length and 0.08 mm in diameter) to the mixing chamber. These conditions were satisfied between 0.05 and 0.5 Hz. The data presented here were obtained at 0.1 Hz.

Solid samples were grown by using the blocked capillary method. In addition to forming samples from ^{4}He gas containing 1 p.p.b. or 0.3 p.p.m. ^{3}He, dilute solid mixtures were grown with ^{3}He
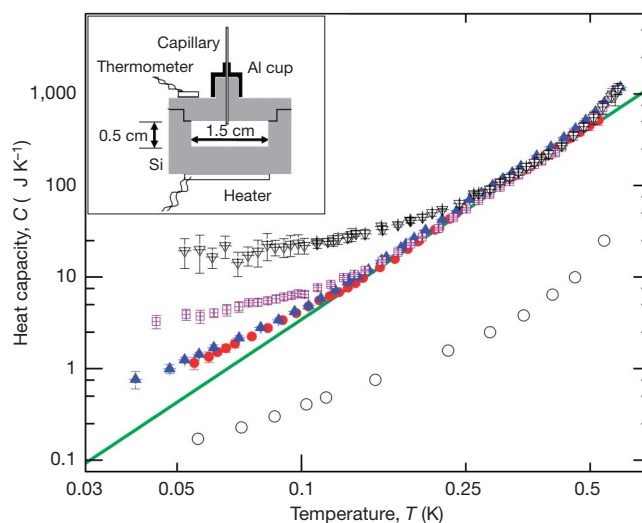

Figure 1 | Heat capacity of four samples containing different amounts of ^{3}He impurities. The green line is the expected T^3 dependence for a Debye solid. The error bar at each T is determined by the s.d. of 10–20 successive measurements, and is typically 2% of C. The measured temperature oscillation scales inversely with the heat capacity; the scatter at low temperature is the largest for the 30 p.p.m. sample. Black open triangles, 30 p.p.m.; purple open squares, 10 p.p.m.; blue triangles, 0.3 p.p.m.; red circles, 1 p.p.b. The empty-cell background, which has been subtracted, is also shown (black open circles). Inset, diagram of the silicon cell, with an internal volume of 0.926 cm^3.

[1]Department of Physics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA.

concentrations of $x_3 = 10$ p.p.m. and 30 p.p.m. These samples were prepared by condensing a small, measured quantity of ^{3}He into the cell before the introduction of commercially pure ^{4}He. The uncertainties in x_3 of the 10 p.p.m. and 30 p.p.m. samples are estimated to be 30% and 20%, respectively. (See Supplementary Discussion for further information on ^{3}He impurities.) The limited accuracy stems from the extra volume (2 cm^3) with a large surface area that is connected to the sample cell. Because the thermal link between the sample cell and mixing chamber is weak, it takes more than 20 h to grow a solid sample, over which time considerable annealing takes place. Thus, as might be expected, we have found additional annealing of samples to have no effect on the measured heat capacity.

In Fig. 1 we show the heat capacity of the empty cell and solid ^{4}He samples with $x_3 = 1$ p.p.b., 0.3 p.p.m., 10 p.p.m. and 30 p.p.m. The contribution of the empty cell, which is one-tenth that of the solid helium over the entire temperature range, has been subtracted from the data. The departure from a T^3 dependence above 500 mK has been observed previously[12–16]. (See Supplementary Discussion for further details.) The most striking finding of this study is the excess heat capacity in addition to the Debye term at low temperature. It is nearly identical for the 1 p.p.b. and 0.3 p.p.m. samples, and becomes discernible below 140 mK. For the 10 p.p.m. and 30 p.p.m. samples the excess term is much larger and shows up below 200 and 250 mK, respectively.

It has been suggested that the observed NCRI is the consequence of solid helium's forming a glassy state and/or the inclusion of glassy regions in the sample[8,9,19,20]. A specific heat that is linear in T is a characteristic of glasses. In an attempt to extract such a linear term, in Fig. 2 we have plotted the temperature dependence of the specific heat C_n (per mole of ^{4}He), divided by T. If, in addition to the T^3 term, there is a component that scales linearly with T, the data will lie on a straight line with a non-zero y intercept. The data for $x_3 = 1$ p.p.b. and 0.3 p.p.m. between 140 and 430 mK do yield straight lines, but both lines extrapolate exactly through the origin, indicating that the specific heat above 140 mK is purely T^3. For comparison, the linear term in the specific heat deduced from recent pressure measurements[20] between 100 and 500 mK has a value of

$\sim$3 mJ mol^{-1} K^{-2}. This term was found to decrease by an order of magnitude after annealing of the solid sample. Given the scatter in the pressure data[20], the reduced value is indistinguishable from zero. We also searched for, but failed to find, any hysteresis or time-dependent effects characteristic of a glassy system. These results suggest that the glassy state is not present in solid samples that are cooled gradually or that have been annealed.

The data in Fig. 2 for the 10 p.p.m. and 30 p.p.m. samples do not fall on straight lines. Instead, the data seem to diverge in the low-temperature limit. This behaviour, which is more prominent in the 30 p.p.m. sample, suggests the presence of a temperature-independent term in addition to the Debye contribution. To test this idea, we plot C_n against T^3 in Fig. 3. In view of Fig. 2, it is not surprising that the data for $x_3 = 1$ p.p.b. and 0.3 p.p.m. fall on a straight line passing through the origin. The fact that the data for the 10 p.p.m. and 30 p.p.m. samples also fall on straight lines with non-zero intercepts confirms our conjecture of a constant term in the specific heat. When we fit each data set in the range 140 mK $< T <$ 430 mK, we obtain Debye temperatures (in K) of $\Theta = 29.1$, 27.6, 29.6 and 29.5 for the 1 p.p.b., 0.3 p.p.m., 10 p.p.m. and 30 p.p.m. samples, respectively. These values are within 10% of the expected values at similar pressures ($\sim$33 bar), with the main uncertainty coming from the absolute measurement of density. There is no *in situ* pressure gauge on the sample cell, but the initial pressure P_i and temperature T_i of the liquid before the formation of the solid block, as well as the approximate freezing point (breakaway point from the melting curve), are known. The densities deduced from P_i and T_i (0.192 $\pm$ 0.01 g cm^{-3}) are consistent with the observed freezing temperatures ($\sim$1.8 $\pm$ 0.1 K, corresponding to 0.194 $\pm$ 0.02 g cm^{-3}). The constant terms from the fit for $x_3 = 10$ p.p.m. and 30 p.p.m. are 59 $\pm$ 10 and 430 $\pm$ 40 μJ mol^{-1} K^{-1}, respectively. The y intercepts for the 1 p.p.b. and 0.3 p.p.m. samples, as shown in the inset to Fig. 3, are indistinguishable from zero; that is, they are within $\pm$5 μJ mol^{-1} K^{-1}. The slope (and thus Θ) and intercept from each fit change insignificantly if all the data below 430 mK are included in the analysis. The likely origin of the constant term is the presence of the ^{3}He impurities in the solid, which for $x_3 = 10$ p.p.m. and 30 p.p.m. correspond to 0.7 $\pm$ 0.2k_B and 1.7 $\pm$ 0.3k_B (where k_B is
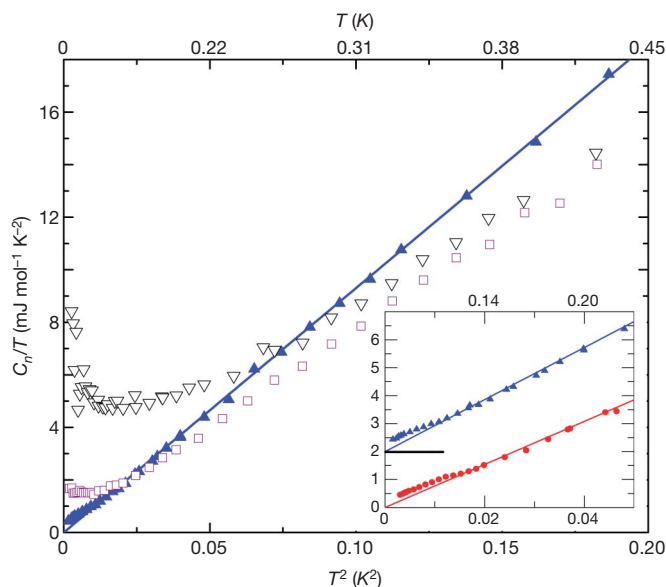

Figure 2 | Plot of C_n/T against T^2 for $x_3 = 1$ p.p.b., 0.3 p.p.m., 10 p.p.m. and 30 p.p.m. Black open triangles, 30 p.p.m.; purple open squares, 10 p.p.m.; blue triangles, 0.3 p.p.m.; red circles, 1 p.p.b. For clarity, the 1 p.p.b. data set is shown only in the inset. The blue line is the linear fit to the 0.3 p.p.m. data in the range 140 mK $< T <$ 430 mK. Inset, demonstration that the deviation from T^3 seen in the 1 p.p.b. and 0.3 p.p.m. samples exists only below 140 mK. In the inset the 0.3 p.p.m. data have been shifted upwards by 2 mJ mol^{-1} K^{-2}.
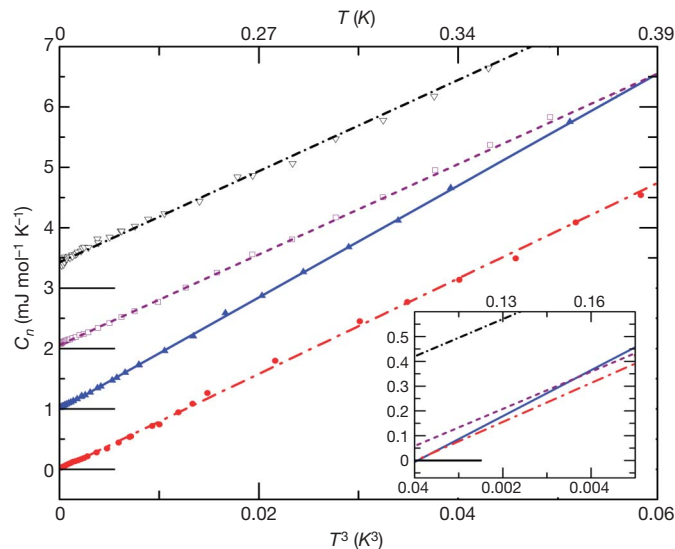

Figure 3 | Plot of specific heat against T^3. Data for the 0.3 p.p.m. (blue triangles), 10 p.p.m. (purple open squares) and 30 p.p.m. (black open triangles) samples have been shifted upwards by 1, 2 and 3 mJ mol^{-1} K^{-1}, respectively. The straight lines are linear fits of C_n versus T^3 between 140 mK and 430 mK. The deviation below 140 mK seen in the 1 p.p.b. (red circles) and 0.3 p.p.m. samples does not significantly affect the straight-line fit. Inset, fits without data points to demonstrate the zero intercepts for $x_3 = 1$ p.p.b. and 0.3 p.p.m. and the non-zero values for $x_3 = 10$ p.p.m. and 30 p.p.m.

the Boltzmann constant) per ^{3}He atom, respectively. There may also be a constant specific heat of the same order (per ^{3}He atom) for the 0.3 p.p.m. and 1 p.p.b. samples but it is undetectable given the small x_3.

NMR measurements[21–23] of dilute ^{3}He in solid ^{4}He have found that below 1.5 K the spin diffusion coefficient D increases with decreasing temperature, until it becomes independent of temperature below 0.8 K. For 100 p.p.m. $< x_3 <$ 500 p.p.m., and between 0.5 and 0.8 K, the diffusion coefficient satisfies[23] the relation $Dx_3 = 2.8 \times 10^{-11}$ cm^2. It is reasonable to assume that D continues to increase for even lower x_3. This behaviour has been interpreted in terms of an impuriton model in which ^{3}He impurities propagate coherently at low temperature as a wavelike excitation[21]. The constant specific heat found in the 10 p.p.m. and 30 p.p.m. samples is probably related to the high mobility of the ^{3}He impurities. (See Supplementary Discussion and Supplementary Fig. 1 for further details on ^{3}He impurities.) It is interesting that the magnitude ($\sim$1 k_B per ^{3}He atom) of the temperature-independent specific heat for $x_3 = 10$ p.p.m. and 30 p.p.m. is on the same order as that of a classical gas. The Fermi degeneracy temperatures of the 10 p.p.m. and 30 p.p.m. samples are 3 mK and 6 mK, respectively, well below the minimum temperature of the experiment.

When the T^3 phonon contribution is subtracted, a broad peak in the specific heat, C_{peak}, centred near 75 mK, is found for the 1 p.p.b., 0.3 p.p.m. and 10 p.p.m. samples. This is shown in the inset to Fig. 4. The peak for the 10 p.p.m. sample rides on top of the constant term of 59 μJ mol^{-1} K^{-1} deduced from Fig. 3. The main panel of Fig. 4 shows the peak with this constant term subtracted, together with those of the 1 p.p.b. and 0.3 p.p.m. samples. The data at high temperature ($T > 150$ mK) have larger scatter because the uncertainty scales with the total measured heat capacity. For the same reason the possible peak is indiscernible for $x_3 = 30$ p.p.m. The peak height in all three samples shown in Fig. 4 is $\sim$20 μJ mol^{-1} K^{-1} ($\sim$2.5 $\times$ 10^{-6} k_B per ^{4}He atom). The peak cannot be related to phase separation of the ^{3}He-^{4}He mixtures[17,24,25], in that it is not hysteretic and its magnitude is completely insensitive to a variation in x_3 of four orders of magnitude. (See Supplementary Discussion for a more detailed discussion on phase separation.) If we assume the specific heat extrapolates

linearly to zero at $T = 0$ K, the additional entropy associated with the peak increases from zero to about 28 μJ mol^{-1} K^{-1} (3.5 $\times$ 10^{-6} k_B per ^{4}He atom) as the temperature is raised from 0 K to 140 mK.

The melting curve of pure ^{4}He was recently measured[26,27] at high resolution between 10 and 320 mK. After the data below 80 mK had been adjusted for an instrumental effect[27], the melting curve followed a T^4 dependence with an uncertainty of about 1 μbar. This implies a T^3 specific heat for both the solid and liquid phases. If we assume that solid in coexistence with liquid is thermodynamically identical to that near 33 bar, the additional contribution to the pressure that stems from the peak in Fig. 4 would cause a deviation of about 5 μbar in a T^4 power-law fit of the melting curve. We do not have an explanation for this apparent discrepancy.

As noted above, the onset of NCRI in 1 p.p.b. solid helium samples is 75 mK (ref. 10). There is evidence that the higher T_o values found in 0.3 p.p.m.[1–6,8,10] samples and at larger concentrations (E. Kim, J. S. Xia, J. T. West, X.L. and M.H.W.C., unpublished observations) are due to a finite vortex response time[6,11], and that the 'true' transition temperature in the zero frequency limit may also be near 75 mK, independently of ^{3}He concentration. A systematic study of the frequency dependence of NCRI in samples of different x_3 will be needed to test this speculation. Nevertheless, the coincident peak in heat capacity and onset of NCRI in 1 p.p.b. samples make it probable that we have observed the thermodynamic signature related to NCRI. The temperature dependence of NCRI in 1 p.p.b. samples[10] may be consistent with a transition that falls into the same universality class as the lambda transition in liquid ^{4}He. If this is indeed so, NCRIF can be calculated by means of the two-scale-factor universality hypothesis. This prediction has been tested and found satisfactory for the superfluid transition of liquid ^{4}He under pressure[28] and for liquid helium confined in porous media[29,30]. If we use the peak height as the amplitude of the singular specific heat then, when compared with the superfluid transition, the NCRIF should be 0.06%. This number lies between 0.03% (E. Kim, J. S. Xia, J. T. West, X.L. and M.H.W.C., unpublished observations) and 0.3% (ref. 10), which are the fractions observed in torsional oscillator studies of 1 p.p.b. samples. This comparison indicates the magnitude of the specific heat peak is at least not unphysical.

Thus, our measurements of specific heat in solid ^{4}He with 10 p.p.m. and 30 p.p.m. ^{3}He found evidence of a constant-specific-heat term that scales with ^{3}He concentration. This term is probably related to the high, temperature-independent, mobility of dilute ^{3}He impurities found in NMR studies. We also found a broad peak in specific heat centred near 75 mK in 1 p.p.b., 0.3 p.p.m. and 10 p.p.m. samples. This peak is very likely to be the thermodynamic signature of the supersolid phase.
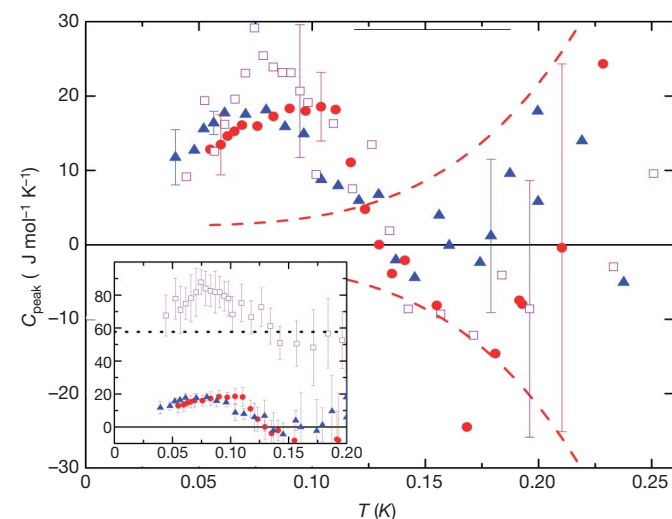
Figure 4 | Specific heat peaks of the 1 p.p.b., 0.3 p.p.m. and 10 p.p.m. samples. Purple open squares, 10 p.p.m.; blue triangles, 0.3 p.p.m.; red circles, 1 p.p.b. The x_3-independent peak centred near 75 mK is revealed when the phonon contribution is subtracted. The s.d. of the 1 p.p.b. data set is shown as red dashed lines. The s.d. for $x_3 = 0.3$ p.p.m. is comparable. For $x_3 = 10$ p.p.m., it is similar above 200 mK but decreases more gradually with decreasing temperature. Inset, comparison of the specific heat of the three samples before subtraction of the constant term of the 10 p.p.m. sample (dotted line at 59 μJ mol^{-1} K^{-1}).

1. Kim, E. & Chan, M. H. W. Probable observation of a supersolid helium phase. *Nature* **427**, 225–227 (2004).
2. Kim, E. & Chan, M. H. W. Observation of superflow in solid helium. *Science* **305**, 1941–1944 (2004).
3. Rittner, A. S. C. & Reppy, J. D. Observation of classical rotational inertia and nonclassical supersolid signals in solid ^{4}He below 250 mK. *Phys. Rev. Lett.* **97**, 165301 (2006).
4. Kondo, M., Takada, S., Shibayama, Y. & Shirahama, K. Observation of non-classical rotational inertia in bulk solid ^{4}He. *J. Low Temp. Phys.* **148**, 695–699 (2007).
5. Penzev, A., Yasuta, Y. & Kubota, M. Annealing effect for supersolid fraction in ^{4}He. *J. Low Temp. Phys.* **148**, 677–681 (2007).
6. Aoki, Y., Graves, J. C. & Kojima, H. Oscillation frequency dependence of nonclassical rotation inertia of solid ^{4}He. *Phys. Rev. Lett.* **99**, 015301 (2007).
7. Leggett, A. J. Can a solid be 'superfluid'? *Phys. Rev. Lett.* **25**, 1543–1546 (1970).
8. Rittner, A. S. C. & Reppy, J. D. Disorder and the supersolid state of solid ^{4}He. *Phys. Rev. Lett.* **98**, 175302 (2007).
9. Prokof'ev, N. V. What makes a crystal supersolid? *Adv. Phys.* **56**, 381–402 (2007).
10. Clark, A. C., West, J. T. & Chan, M. H. W. Nonclassical rotational inertia in helium crystal. *Phys. Rev. Lett.* **99**, 135302 (2007).
11. Anderson, P. W. Two new vortex liquids. *Nature Phys.* **3**, 160–162 (2007).
12. Heltemes, E. C. & Swenson, C. A. Heat capacity of solid ^{3}He. *Phys. Rev.* **128**, 1512–1519 (1962).

13. Edwards, D. O. & Pandorf, R. C. Heat capacity and other properties of hexagonal close-packed helium-4. *Phys. Rev.* **140,** A816–A825 (1965).

14. Gardner, W. R., Hoffer, J. K. & Phillips, N. E. Thermodynamic properties of ^{4}He. The hcp phase at low densities. *Phys. Rev. A.* **7,** 1029–1043 (1973).

15. Castles, S. H. & Adams, E. D. Specific heat of solid helium. *J. Low Temp. Phys.* **19,** 397–431 (1975).

16. Hébral, B., Frossati, G., Godfrin, H., Thoulouze, D., Greenberg, A. S. & in. *Phonons in Condensed Matter* (ed. Maris, H. J.) 169–172 (Plenum, New York, 1980).

17. Clark, A. C. & Chan, M. H. W. Specific heat of solid helium. *J. Low Temp. Phys.* **138,** 853–858 (2005).

18. Sullivan, P. F. & Seidel, G. Steady-state, ac-temperature calorimetry. *Phys. Rev.* **173,** 679–685 (1968).

19. Nussinov, Z., Balatsky, A. V., Graf, M. J. & Trugman, S. A. Origin of the decrease in the torsional-oscillator period of solid ^{4}He. *Phys. Rev. B* **76,** 014530 (2007).

20. Grigorev, V. N. *et al.* Observation of a glassy phase in solid ^{4}He in the supersolidity region. Preprint at ⟨http://arxiv.org/abs/cond-mat/0702133⟩ (2007).

21. Richards, M. G., Pope, J. & Widom, A. Evidence for isotopic impuritons in solid helium. *Phys. Rev. Lett.* **29,** 708–711 (1972).

22. Grigor'ev, V. N. *et al.* ^{3}He impurity excitations in solid ^{4}He. *J. Low Temp. Phys.* **13,** 65–79 (1973).

23. Allen, A. R., Richards, M. G. & Schratter, J. Anomalous temperature dependence of *D* and T_2 for dilute solutions of ^{3}He in solid ^{4}He. *J. Low Temp. Phys.* **47,** 289–320 (1982).

24. Hébral, B. *et al.* Fermi-liquid droplets in liquid-solid solutions of the helium isotopes. *Phys. Rev. Lett.* **46,** 42–45 (1981).

25. Schrenk, R., Friz, O., Fujii, Y., Syskakis, E. & Pobell, F. Heat capacity and pressure at phase separation and solidification of ^{3}He in hcp ^{4}He. *J. Low Temp. Phys.* **84,** 133–156 (1991).

26. Todoshchenko, I. A. *et al.* Melting curve of ^{4}He: no sign of a supersolid transition down to 10 mK. *Phys. Rev. Lett.* **97,** 165302 (2006).

27. Todoshchenko, I. A., Alles, H., Junes, H. J., Parshin, A., Ya. & Tsepelin, V. Absence of low temperature anomaly on the melting curve of ^{4}He. *JETP Lett.* **85,** 555–558 (2007).

28. Singsaas, A. & Ahlers, G. Universality of static properties near the superfluid transition in ^{4}He. *Phys. Rev. B* **30,** 5103–5115 (1984).

29. Yoon, J. & Chan, M. H. W. Superfluid transition of ^{4}He in porous gold. *Phys. Rev. Lett.* **78,** 4801–4804 (1997).

30. Zassenhaus, G. M. & Reppy, J. D. Lambda point in the ^{4}He-Vycor system: a test of hyperuniversality. *Phys. Rev. Lett.* **83,** 4800–4803 (1999).

Author Contributions X.L., A.C.C. and M.H.W.C. contributed equally to this work.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.H.W.C. (chan@phys.psu.edu).

LETTERS

Attosecond spectroscopy in condensed matter

A. L. Cavalieri[1], N. Müller[2], Th. Uphues[1,2], V. S. Yakovlev[3], A. Baltuška[1,4], B. Horvath[1], B. Schmidt[5], L. Blümel[5], R. Holzwarth[5], S. Hendel[2], M. Drescher[6], U. Kleineberg[3], P. M. Echenique[7], R. Kienberger[1], F. Krausz[1,3] & U. Heinzmann[2]

Comprehensive knowledge of the dynamic behaviour of electrons in condensed-matter systems is pertinent to the development of many modern technologies, such as semiconductor and molecular electronics, optoelectronics, information processing and photovoltaics. Yet it remains challenging to probe electronic processes, many of which take place in the attosecond (1 as = 10^{-18} s) regime. In contrast, atomic motion occurs on the femtosecond (1 fs = 10^{-15} s) timescale and has been mapped in solids in real time[1,2] using femtosecond X-ray sources[3]. Here we extend the attosecond techniques[4,5] previously used to study isolated atoms in the gas phase to observe electron motion in condensed-matter systems and on surfaces in real time. We demonstrate our ability to obtain direct time-domain access to charge dynamics with attosecond resolution by probing photoelectron emission from single-crystal tungsten. Our data reveal a delay of approximately 100 attoseconds between the emission of photoelectrons that originate from localized core states of the metal, and those that are freed from delocalized conduction-band states. These results illustrate that attosecond metrology constitutes a powerful tool for exploring not only gas-phase systems, but also fundamental electronic processes occurring on the attosecond timescale in condensed-matter systems and on surfaces.

Photoemission spectroscopy is based on the photoelectric effect, first explained by Einstein more than 100 years ago[6]. According to Einstein's law, photoelectrons ejected from a metal surface by light will have a kinetic energy that depends on the incident photon energy and the electron's original bound state energy. Photoelectron spectra will thus provide information about the electronic structure of the metal if well-characterized light sources are used[7]. Indeed, experiments using a broad range of photon energies have now been performed to determine the steady-state electronic properties of many bulk materials, thin films, and surfaces. The photoemission process itself involves three steps: excitation, transport, and ultimately escape of the photoelectron through the surface[8]. Here, in a proof-of-principle experiment, we combine this spectroscopy with attosecond temporal resolution to obtain time-domain insight into the electron transport stage of the photoemission process. The measurements represent (to our knowledge) the first direct attosecond time-resolved observation of electron transport in a condensed-matter system, and we expect that they will trigger other experimental research into the dynamics of processes that have attracted interest in solid-state and surface science. Such processes include charge transfer[9,10], charge screening[11], image charge creation and decay[12], electron–electron scattering[13], and collective electronic motion[14].

Time-resolved photoemission spectroscopy was originally implemented in the picosecond (1 ps = 10^{-12} s) to femtosecond regime,

using first visible[15–17] and then extreme ultraviolet (XUV)[18,19] radiation. These experiments utilize one light pulse to trigger the dynamics, followed by a second light pulse to induce photoemission and thereby probe the transient state. Experiments using the laser-assisted photoelectric effect have been carried out[20–22]; but the XUV photoemission lasted over several wave cycles of the coincident near-infrared (NIR) light, limiting the time resolution to >10 fs. To overcome this limitation, we use single sub-femtosecond XUV pulses[4,5] for pumping, and coincident NIR waveform-controlled few-cycle laser pulses[23] as a probe[5]. The XUV pulse triggers the photoemission process, with only those photoelectron wave packets initiated from the uppermost atomic layers escaping without inelastic collision.
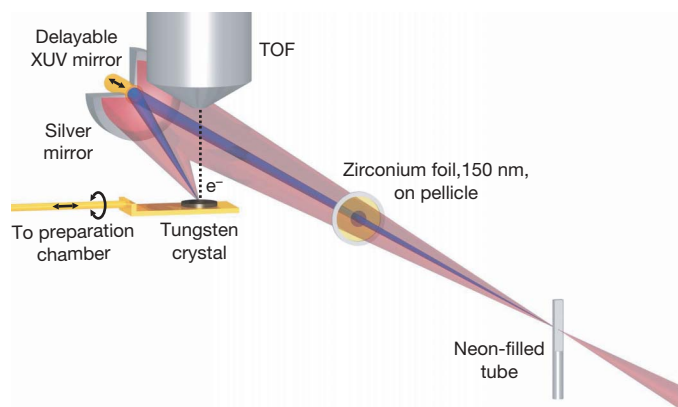

Figure 1 | Experimental set-up. Waveform-controlled, ~5-fs, 750-nm, 400-μJ laser pulses are focused with a mirror of 500-mm focal length into a ~2-mm-diameter tube filled with neon to generate XUV radiation by high-harmonic generation. The collinear XUV and NIR beams co-propagate towards a cored two-part mirror in the measurement chamber that is maintained under ultrahigh vacuum in a 1-m-length differential vacuum pumping stage, until the beams are separated by a pellicle/zirconium foil assembly. The zirconium foil transmits the XUV but blocks the NIR. The XUV radiation, indicated by the blue beam, is incident on a 6 eV (full-width at half-maximum FWHM) broad multilayer band-pass reflector centred at ~91 eV, which is mounted on a piezo-electric delay stage. With a proper XUV spectrum, the multilayer mirror reflects and focuses ~300-as (FWHM) XUV pulses. The NIR pulse, indicated by the violet beam, is reflected by a stationary (silver) outer annular mirror confocal with the inner mirror (*f* = 12.5 cm). Both pulses are focused onto the (110) surface of a tungsten single crystal that is mounted on a manipulator to control the angle of incidence. The manipulator is also used to retract the crystal into a preparatory chamber for cleaning. Resultant XUV-induced photoemission, which is detected by the time-of-flight spectrometer (TOF), is streaked by the coincident NIR laser pulse.

[1]Max-Planck-Institut für Quantenoptik, Hans-Kopfermann-Str. 1, D-85748 Garching, Germany. [2]Fakultät für Physik, Universität Bielefeld, D-33615 Bielefeld, Germany. [3]Department für Physik, Ludwig-Maximilians-Universität, Am Coulombwall 1, D-85748 Garching, Germany. [4]Institut für Photonik, Technische Universität Wien, Gußhausstr. 27, A-1040 Wien, Austria. [5]Menlo Systems GmbH, Am Klopferspitz 19, D-82152 Martinsried, Germany. [6]Institut für Experimentalphysik, Universität Hamburg, Luruper Chaussee 149, D-22761 Hamburg, Germany. [7]Dpto. Fisica de Materiales UPV/EHU, Centro Mixto CSIC-UPV/EHU and Donostia International Physics Center (DPIC), Paseo Manual de Lardizabal 4, 20018 San Sebastian, Spain.

Photoexcited electron wave packets propagate through the material in upper conduction bands, ultimately leaving the surface with an average kinetic energy determined by the XUV photon energy, the initial binding energy, and the material work function. An attosecond transient recorder (ATR), previously developed and used in gas-phase experiments[5], is used to observe the emitted photoelectron wave packet. In this scheme, the photoelectron momentum is further influenced by the electric field of a coincident few-cycle NIR laser pulse, giving rise to a 'streaked' final momentum distribution[24,25].

An ATR spectrogram is compiled by measuring a series of streaked photoelectron spectra with a time-of-flight detector, recorded as a function of time delay between the XUV pump and NIR streaking field. Important characteristics of the emitted electron wave packets, including their duration and frequency sweep, or 'chirp', can be determined from the spectra[24,25]. If measured for two or more different types of electrons, the complete ATR spectrograms can also yield relative timing information about the arrival of the wave packets on the surface, because the streaking effect is negligible until the electrons emerge from the surface (see Methods). The resolution of the ATR depends on the duration of the XUV excitation, the gradient of the streaking NIR field, and the signal-to-noise ratio in the photoelectron spectra.

In comparison to experiments performed on isolated atoms, ATR measurements in condensed-matter systems are more complicated because the photoelectron wave packets can be released from energy bands containing many distinct states rather than from a single, isolated energy level. Unoccupied conduction-band states just above the Fermi energy (defined by the highest occupied energy level in the absence of thermal excitation) might become populated by single-photon absorption of the leading edge of the NIR probe field prior to XUV photoemission, which is an unwanted complication. In contrast to conduction-band states, the localized 4f core states of tungsten are deeply bound and fully populated. Therefore, these states are unsusceptible to this potential influence of the streaking field, and constitute an ideal test case for proof of the extension of attosecond metrology to solids.

As a further challenge, above-threshold ionization (ATI) by the streaking field can, in condensed matter, generate energetic photoelectrons, obscuring detection of XUV-induced photoelectrons. ATI is favoured by the low work function of metals (as compared to the relatively high ionization potential of isolated atoms), which limits the intensity of the applied streaking field to levels far below those that can be used in gas-phase experiments.

As indicated by the diagram of the experimental set-up in Fig. 1, streaked photoemission spectra from a tungsten (110) crystal surface were recorded by collecting electrons within a narrow cone aligned perpendicularly to the surface. The relative delay between the XUV pulse and the NIR waveform-controlled streaking field was varied in 300-as steps in a sequence chosen to minimize systematic error, with
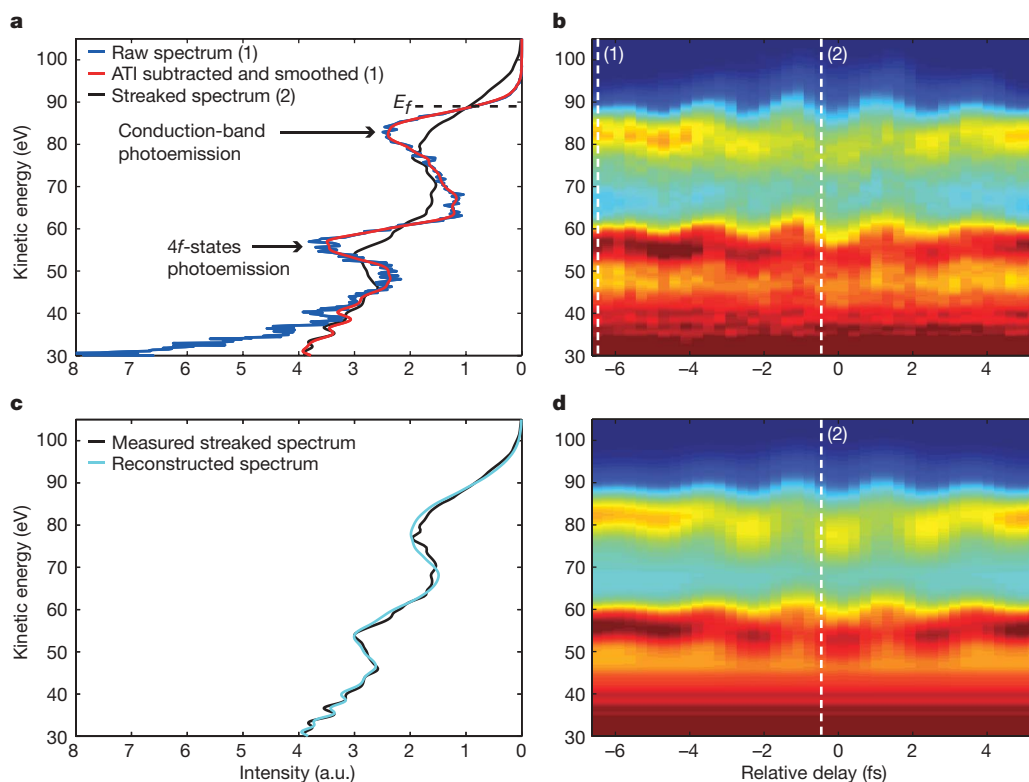

Figure 2 | Attosecond time-resolved photoemission spectra.
a, Photoelectron spectra collected at two different relative delays. The spectrum recorded at the delay indicated by the white dashed line labelled '(1)' in **b** is far from the zero of delay as defined by the overlap of the maximums of the NIR and XUV pulse envelopes. This spectrum shows pronounced peaks corresponding to the 4f-state and conduction-band photoemission. The blue line shows the raw spectrum as recorded by the time-of-flight detector, and the red line shows the corresponding spectrum after subtraction of NIR-induced ATI background and numerical smoothing. The 4f photoemission is peaked near 56 eV. The conduction-band photoemission is peaked near 83 eV, owing to a high density of d-band conducting states just below the Fermi energy. E_f denotes the kinetic energy of a photoelectron excited from the Fermi energy level. The other spectrum in **a** (displayed only after ATI subtraction and smoothing) was recorded near zero delay, as indicated by the white dashed line labelled '(2)'. At this delay,

the XUV pulse peak coincides with the NIR field maximum on our target. Consequently, the NIR vector potential crosses zero, giving rise to the strongest streaking and a clear broadening of the photoemission peaks is observed. **b,** The full ATI subtracted spectrogram for both the 4f-states and conduction-band photoemission. The streaking waveform (vector potential) is evident in both spectrograms, proving the extension of attosecond metrology to condensed-matter systems. We can expect that the energy modulation of the conduction-band peak should be a little larger than that of the 4f peak owing to their initial photoelectron kinetic energy. The amplitude of the spectral shift has a square-root dependence on initial kinetic energy (KE)[24], yielding an expected ratio in spectral shift of $\sqrt{KE_{cond}}\big/\sqrt{KE_{4f}} \approx 1.2$, which is consistent with our measurements. **c,** The reconstructed spectrum corresponding to that measured and displayed in **a. d,** The full simulated spectrogram.

spectra integrated for 60 s at each delay. (For a brief description of the experiment, see Methods Summary; full details regarding set-up, measurements and data analysis are provided as Supplementary Information.) Characteristic spectra obtained with our system are shown in Fig. 2a, indicating that emission from the conduction band occurs at a kinetic energy of ~83 eV while emission from the localized 4f states occurs at ~56 eV. At kinetic energies significantly below the 4f peak, the measured spectrum is due to NIR-induced ATI photoelectrons and XUV-generated photoelectrons that have undergone inelastic scattering.

The two distinct background components were distinguished by recording an additional photoelectron spectrum without the NIR streaking field. The ATI component was subsequently subtracted from the measured data (see Supplementary Information). This subtraction is illustrated for a fixed delay in Fig. 2a, and was performed at each of the delay steps, resulting in the full spectrogram presented in Fig. 2b. Here, a positive relative delay corresponds to the XUV pulse arriving earlier with respect to the streaking field at the surface. Both the 4f and conduction-band photoemission exhibit a pronounced periodic upshift and downshift in energy as a function of relative delay and, as in previous gas-phase experiments, the spectrogram reveals the waveform (vector potential) of the streaking field[5,26,27]. Our ability to resolve the field oscillation indicates that the photoemission from the 4f core states and from the conduction band is subfemtosecond in duration, and proves that attosecond metrology has been successfully extended to condensed-matter systems.

Further examination reveals that the 4f spectrogram is shifted along the delay coordinate with respect to the conduction-band spectrogram. This effect is readily apparent upon inspection of the smoothed spectrograms that are obtained by interpolation of the measured data and shown in Fig. 3a. We quantify the temporal shift in the measured data by evaluating, for each delay step, the centre-of-mass (COM) of the spectral regions spanning the 4f and conduction-band peaks that cover the energy intervals 47–66 eV and 66–110 eV, respectively. Characterizing the periodic motion of the peaks through their COM requires no assumptions or fitting parameters, yet yields timing information that is invariant to fluctuations in the instantaneous laser parameters. The approach is also relatively insensitive to inelastic scattered background photoelectrons, which could not be subtracted from our measurements. As a result, the COM accurately describes the streaking-induced time-dependence of the energy shift of the 4f and conduction-band peaks, as shown in Fig. 3b.

By comparing the COM trajectories of the 4f and conduction band at the seven zero-crossings of the vector potential, we obtain seven independent measurements of their relative timing. This yields a temporal shift of $\Delta\tau = 110 \pm 70$ as between the ATR spectrograms of the conduction-band and 4f photoelectrons. (The error estimate results from a straightforward extrapolation of the error in calculating the COM; see Supplementary Information.) This shift or delay was observed in different independent measurements made at different locations on the tungsten sample, with the results corroborating the above value of $\Delta\tau$ to within the measurement error. We note that the rather large error associated with our $\Delta\tau$ value could be most effectively reduced in future measurements by using higher XUV photon energies and fluxes.

The shift between the two spectrograms indicates that, on average, photoelectrons originating from the localized 4f states emerge from the tungsten surface approximately 100 as later than those originating from the delocalized conduction band—even though the photoemission process for both types of electrons is initiated simultaneously by the same XUV pulse. The delay effect thus occurs during transport of the excited photoelectrons to the surface, illustrating that our technique provides a means to directly observe features of electron wave packet propagation towards the surface with attosecond precision.

By adapting a quantum mechanical model used in previous gasphase streaking experiments[28], we are able to reconstruct the measured

spectra and spectrograms. The modelling of the streaking experiment requires some assumptions, leaving several parameters (such as duration of the electron wave packets, their chirp, and their emission time) for optimization. Figure 2c and d shows the reconstructions that best agree with experiment. These were obtained for wave packets with a duration of ~300 as (full-width at half-maximum, FWHM) and assuming a delay of ~100 as between the emission times of the electron wave packets, which supports the conclusions drawn from the COM analysis.

Our measurements also indicate that electron wave packets launched from both the localized 4f and delocalized conduction-band states are nearly undistorted on propagation to the surface. To explain the observed delay, we consider the group velocities for the two different photoelectron wave packets travelling in the solid. The crucial point is that after absorption of an XUV photon, the electron is excited into an upper conduction band region that depends on the electron's
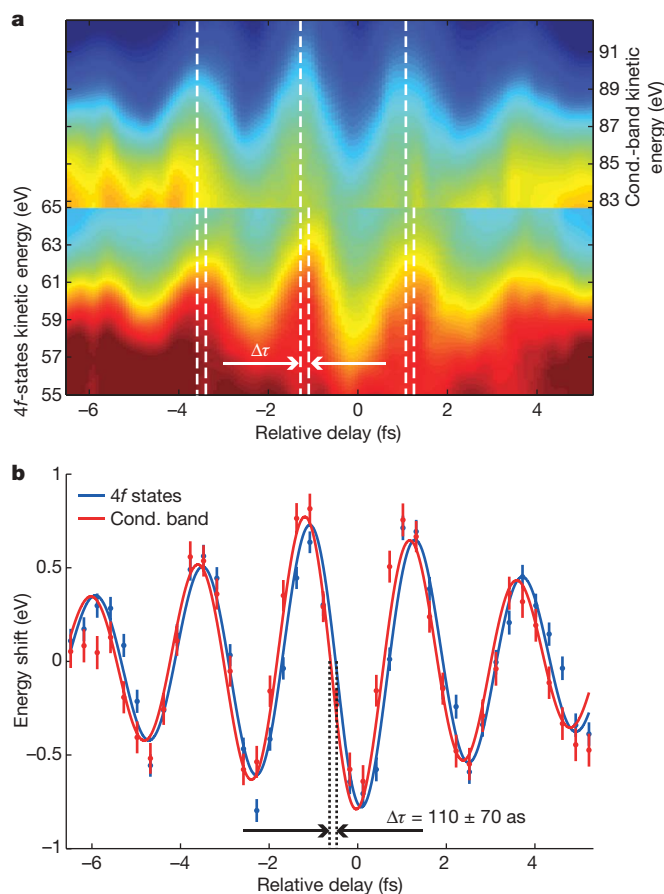

Figure 3 | Evidence of delayed photoemission. a, The 4f and conduction-band spectrograms, following cubic-spline interpolation of the measured data (but without background subtraction). The spectral region between ~65 eV and ~83 eV has been omitted to more easily compare the edges of the 4f and conduction-band peaks. A small shift in the relative delay is evident, as indicated by the white dashed lines through the fringes, and can be seen at each fringe maximum and minimum. Quantification of the shift of the 4f with respect to the conduction-band spectrogram is made by COM analysis, and is summarized in **b**. The energy intervals, within which the COMs were calculated, are 47–66 eV for the 4f photoemission peak and 66–110 eV for the conduction-band photoemission peak. Vertical error bars (±1 s.d.) are calculated from noise in the measured spectra (see Supplementary Information for details). For ease of visual comparison, the COM energy shift of the 4f spectral region was scaled by a factor of 2.5, to offset the stabilizing effect of the background plateau underneath the 4f peak (see Supplementary Information), in order to illuminate the ~100 as delay in emission. Rescaling these COM data points along the energy axis cannot influence the measured delay. The COM data points were fitted with a damped sinusoid, which corresponds to the NIR streaking field, to guide the eye.

initial binding energy and the XUV photon energy. If the relationship between momentum and energy of such an electron were that of a free electron, the ratio of velocities of the conduction-band and 4*f* electrons travelling with energies of 85 eV and 58 eV, respectively, would be a factor of ~1.2 (energy is defined with respect to the Fermi energy inside the material). However, elastic interactions with atoms in the crystal lattice, which give rise to electronic band structure, modify the momentum–energy relationship. This implies that in tungsten for XUV photon energies of ~91 eV, the mean velocity of the conduction-band photoelectrons is approximately twice that of the 4*f* photo-electrons (see Supplementary Information). We also note that due to their longer inelastic mean free paths[29], the slow 4*f* photoelectrons originate, on average, from ~1 Å deeper in the tungsten crystal than the fast conduction-band photoelectrons. On the basis of these considerations, we estimated the absolute delay between the initial excitation of a photoelectron and its escape through the surface to be ~60 as and ~150 as for the conduction and 4*f* photoelectrons, respectively. These values suggest a relative delay between photo-electron emissions from the surface of ~90 as, which is in good agreement with the observed delay in the emission of the 4*f* and conduction-band photoelectrons.

In summary, our observations demonstrate the successful extension of attosecond metrology to condensed-matter systems. Although the current experimental apparatus provides access to only the relative group delay in electron wave packet propagation, future measurements of absolute emission delays may be feasible using the same methods, for example by direct comparison with gas-phase ATR data. At this point, attosecond photoemission spectroscopy presents a clear path toward ultimately uncovering the intermediate processes leading to ejection of a photoelectron. This information might shed new light on data previously obtained with conventional time-integral photoemission spectroscopy and allow for accurate description of charge dynamics on the electronic timescale in both condensed matter and on surfaces.

METHODS SUMMARY

A 1-kHz-repetition-rate, waveform-controlled, few-cycle, ~5-fs, 400-μJ, 750-nm Ti:sapphire laser system is the front-end of our apparatus and is used in combination with proper spectral filtering to efficiently generate isolated attosecond pulses of XUV radiation by high-harmonic generation. As shown in Fig. 1, the XUV and NIR pulses co-propagate towards a two-part focusing mirror at near normal incidence. The inner component is a Mo/Si multilayer mirror and reflects the XUV radiation over a bandwidth of ~6 eV (FWHM) centred at ~91 eV, supporting 300-as transform-limited pulses[30]. The XUV mirror is mounted on a translation stage, providing a precise delay between the XUV pump and the NIR streaking pulse. The temporal resolution that can be achieved in pump-probe experiments using these pulses and this apparatus is expected to be a small fraction of the pulse half-width because XUV pulses generated with similar spectra and multilayer optics have previously been fully characterized and observed to be gaussian[24].

The tungsten surface must be sufficiently free of contamination to minimize photoelectron scattering on emission. Therefore, the measurement chamber is maintained under ultrahigh vacuum conditions with typical background pressures of < 10^{-9} mbar, which suppresses the accumulation of contaminates on the crystal surface to a level permitting a full ATR spectrogram to be recorded without interruption.

In our application of the ATR, detection of electrons occurs in the direction normal to the tungsten crystal surface, and the NIR streaking field is incident on the tungsten (110) crystal surface near Brewster's angle (~75°). For this angle of incidence, owing to the refractive index of tungsten, photoelectron wave packets are not efficiently accelerated in the direction of observation until they emerge from the surface. Even though the streaking field penetrates the tungsten crystal, inside the material the electric field component along the surface normal is weaker by a factor of approximately 16, allowing us to neglect streaking effects until the electron wave packets emerge from the surface. The absence of effective streaking until the photoelectron emerges from the surface is generally the case for solids, allowing us to time processes occurring within the material.

Additional, detailed description of the experimental apparatus, measurement technique, and data analysis is provided in Supplementary Information.

1. Reis, D. A. & Lindenberg, A. M. in *Light Scattering in Solids IX* (eds Cardona, M. & Merlin, R.) 371–422 (Topics in Applied Physics 108, Springer, Berlin, 2007).
2. Fritz, D. M. *et al.* Ultrafast bond softening in bismuth: Mapping a solid's interatomic potential with X-rays. *Science* **315**, 633–636 (2007).
3. Pfeifer, T., Spielmann, C. & Gerber, G. Femto-second X-ray science. *Rep. Prog. Phys.* **69**, 443–505 (2006).
4. Hentschel, M. *et al.* Attosecond metrology. *Nature* **414**, 509–513 (2001).
5. Kienberger, R. *et al.* Atomic transient recorder. *Nature* **427**, 817–821 (2004).
6. Einstein, A. Über einen die Erzeugung und Verwandlung des Lichts betreffenden heuristischen Gesichtspunkt. *Ann. Phys.* **17**, 132–148 (1905).
7. Siegbahn, K. Electron-spectroscopy — outlook. *J. Electron Spectrosc. Relat. Phenom.* **5**, 3–97 (1974).
8. Berglund, C. N. & Spicer, W. E. Photoemission studies of copper and silver: theory. *Phys. Rev.* **136**, A1030–A1044 (1964).
9. Brühwiler, P. A., Karis, O. & Martensson, N. Charge-transfer dynamics studied using resonant core spectroscopies. *Rev. Mod. Phys.* **74**, 703–740 (2002).
10. Föhlisch, A. *et al.* Direct observation of electron dynamics in the attosecond domain. *Nature* **436**, 373–376 (2005).
11. Borisov, A., Sánchez-Portal, D., Díez-Muino, R. & Echenique, P. M. Dimensionality effects in time-dependent screening. *Chem. Phys. Lett.* **387**, 132–137 (2004).
12. Huber, R. *et al.* How many-particle interactions develop after ultrafast excitation of an electron-hole plasma. *Nature* **414**, 286–289 (2001).
13. Haight, R. Electron dynamics at surfaces. *Surf. Sci. Rep.* **21**, 277–325 (1995).
14. Cavalleri, A. *et al.* Tracking the motion of charges in a terahertz light field by femtosecond X-ray diffraction. *Nature* **442**, 664–666 (2006).
15. Yen, R. *et al.* Picosecond laser interaction with metallic zirconium. *Appl. Phys. Lett.* **40**, 185–187 (1982).
16. Höfer, U. *et al.* Time-resolved coherent photoelectron spectroscopy of quantized electronic states on metal surfaces. *Science* **277**, 1480–1482 (1997).
17. Petek, H. & Ogawa, S. Femtosecond time-resolved two-photon photoemission studies of electron dynamics in metals. *Prog. Surf. Sci.* **56**, 239–310 (1997).
18. Haight, R. & Peale, D. R. Tunable photoemission with harmonics of subpicosecond lasers. *Rev. Sci. Instrum.* **65**, 1853–1857 (1994).
19. Siffalovic, P. *et al.* Laser-based apparatus for extended ultraviolet femtosecond time-resolved photoemission spectroscopy. *Rev. Sci. Instrum.* **72**, 30–35 (2001).
20. Schins, J. M. *et al.* Observation of laser-assisted Auger decay in argon. *Phys. Rev. Lett.* **73**, 2180–2183 (1994).
21. Glover, T. E., Schoenlein, R. W., Chin, A. H. & Shank, C. V. Observation of laser assisted photoelectric effect and femtosecond high order harmonic radiation. *Phys. Rev. Lett.* **73**, 2180–2183 (1994).
22. Miaja-Avila, L. *et al.* Laser-assisted photoelectric effect from surfaces. *Phys. Rev. Lett.* **97**, 113604 (2006).
23. Baltuska, A. *et al.* Attosecond control of electronic processes by intense light fields. *Nature* **422**, 611–615 (2003).
24. Quéré, F., Mairesse, Y. & Itatani, J. Temporal characterization of attosecond XUV fields. *J. Mod. Opt.* **52**, 339–360 (2005).
25. Yakovlev, V., Bammer, F. & Scrinzi, A. Attosecond streaking measurements. *J. Mod. Opt.* **52**, 395–410 (2005).
26. Goulielmakis, E. *et al.* Direct measurement of light waves. *Science* **305**, 1267–1269 (2004).
27. Sansone, G. *et al.* Isolated single-cycle attosecond pulses. *Science* **314**, 443–446 (2006).
28. Kitzler, M., Milosevic, N., Scrinzi, A., Krausz, F. & Brabec, T. Quantum theory of attosecond XUV pulse measurement by laser dressed photoionization. *Phys. Rev. Lett.* **88**, 173904 (2002).
29. Tanuma, S., Powell, C. J. & Penn, D. R. Calculations of electron inelastic mean free paths. 2. Data for 27 elements over the 50–2000-eV range. *Surf. Interface Anal.* **17**, 911–926 (1991).
30. Wonisch, A. *et al.* Design, fabrication, and analysis of chirped multilayer mirrors for reflection of extreme-ultraviolet attosecond pulses. *Appl. Opt.* **45**, 4147–4156 (2006).

Clathrate nanostructures for mass spectrometry

Trent R. Northen[1,2]*, Oscar Yanes[1,2]*, Michael T. Northen[4], Dena Marrinucci[3], Winnie Uritboonthai[1,2], Junefredo Apon[1,2], Stephen L. Golledge[5], Anders Nordström[1,2] & Gary Siuzdak[1,2]*

The ability of mass spectrometry to generate intact biomolecular ions efficiently in the gas phase has led to its widespread application in metabolomics[1], proteomics[2], biological imaging[3], biomarker discovery[4] and clinical assays (namely neonatal screens[5]). Matrix-assisted laser desorption/ionization[6,7] (MALDI) and electrospray ionization[8] have been at the forefront of these developments. However, matrix application complicates the use of MALDI for cellular, tissue, biofluid and microarray analysis and can limit the spatial resolution because of the matrix crystal size[9] (typically more than 10 μm), sensitivity and detection of small compounds (less than 500 Da). Secondary-ion mass spectrometry[10] has extremely high lateral resolution (100 nm) and has found biological applications[11,12] although the energetic desorption/ionization is a limitation owing to molecular fragmentation. Here we introduce nanostructure-initiator mass spectrometry (NIMS), a tool for spatially defined mass analysis. NIMS uses 'initiator' molecules trapped in nanostructured surfaces or 'clathrates' to release and ionize intact molecules adsorbed on the surface. This surface responds to both ion and laser irradiation. The lateral resolution (ion-NIMS about 150 nm), sensitivity, matrix-free and reduced fragmentation of NIMS allows direct characterization of peptide microarrays, direct mass analysis of single cells, tissue imaging, and direct characterization of blood and urine.

A diagram of the main applications of NIMS is provided in Supplementary Fig. 1. Conceptually, NIMS reverses the system of desorption/ionization developed by Tanaka[7], which uses ~30-nm nanoparticles suspended in glycerol, where desorption/ionization is thought to occur by a thermally driven process for nanoparticles larger than 3 nm (ref. 13). As illustrated in Fig. 1a, b, NIMS uses a nanostructured surface composed of roughly 10-nm pores (Fig. 1c) to trap 'initiator' materials such as fluorinated siloxanes. The analyte is adsorbed on top of the NIMS surface and laser irradiation results in rapid surface heating causing initiator vaporization from many clathrates, triggering the desorption of adsorbed materials. Scanning electron microscopy (SEM) images of a laser-irradiated NIMS surface suggest that rapid surface heating results in a violent expansion of initiator, as indicated by bulk surface destruction and deformation (Fig. 1a). Ion irradiation also results in intact ion formation; however, SEM imaging (not shown) of ion-irradiated surfaces does not reveal any changes in surface nanostructure, indicating a more localized process. The large amounts of trapped initiators are found to migrate reversibly into and out of the NIMS surface with slight heating and cooling (Supplementary Movie 1), indicating that the initiator is highly ordered within the nanostructured surface. In these ways NIMS is very different from existing surface desorption/ionization techniques such as desorption–ionization on silicon[14] (DIOS), which is limited to small molecules and peptides, and laser

irradiation. In addition, NIMS is characterized by robustness, reproducibility, ease of preparation and minimal training for newcomers (Supplementary Methods), all of which have limited the application of DIOS.

NIMS is flexible as regards irradiation source, surface, and initiator composition: NIMS activity has been found to be compatible with a variety of irradiation including a nitrogen laser and ion sources (Au^+, Ga^+, Bi^+ and Bi_3^+). Native surfaces, silanized surfaces, and even surfaces sputtered with Au/Pd have NIMS activity. A wide range of initiators including lauric acid, polysiloxanes, siloxanes and silanes (molecular masses from 200 to 14,000 Da) have been used for NIMS (Supplementary Table 1), although the use of perfluorinated siloxanes was preferred for laser-NIMS because they are effectively trapped within the nanostructured surface and have the best performance, presumably as a result of the low solubility of analyte within the resulting clathrates. Because of the higher vacuum used
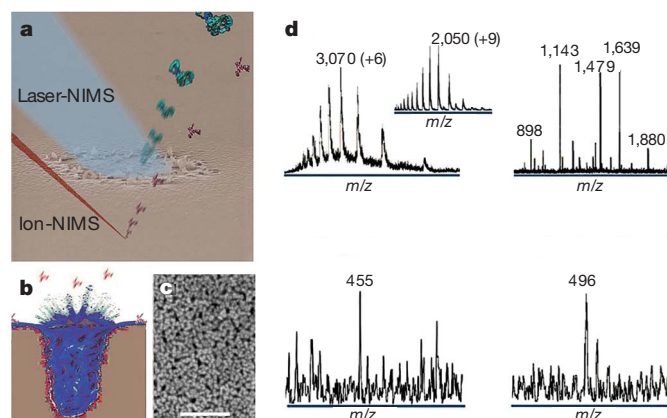

Figure 1 | Nanostructure-initiator mass spectrometry. a, Illustration superimposed on an SEM image of a NIMS surface after irradiation with a single laser shot (blue), revealing localized surface distortion and destruction. By comparison, ion irradiation (red) allows a much higher lateral resolution. **b,** Illustration of possible mechanism in which surface irradiation results in the vaporization or fragmentation of initiator (blue) trapped in a surface pore, triggering analyte desorption/ionization. **c,** SEM image revealing that the NIMS surface is composed of 10-nm pores. Scale bar, 100 nm. **d,** Laser irradiation (wavelength 337 nm) of a NIMS surface. Upper left panel: detection of a multiply charged protein (50 nmol of β-lactoglobulin) in a similar manner to ESI (inset). Upper right panel: detection of a BSA tryptic digest (500 amol). Lower left panel: detection of the calcium antagonist verapamil (700 ymol). Lower right panel: detection of the endogenous metabolite 1-palmitoyllysophosphatidylcholine (50 amol). The initiator was bis(tridecafluoro-1,1,2,2-tetrahydrooctyl)tetramethyl-disiloxane; 0.5-μl drops were used. MALDI and ESI spectra for these samples are provided in Supplementary Figs 2, 3 and 5.

[1]Department of Molecular Biology, [2]Scripps Center for Mass Spectrometry, [3]Department of Cell Biology, The Scripps Research Institute, La Jolla, California 92037, USA. [4]Materials Department, University of California at Santa Barbara, Santa Barbara, California 93106, USA. [5]CAMCOR Surface Analysis Facility at the University of Oregon, Eugene, Oregon 97403, USA.
*These authors contributed equally to this work.

in time-of-flight–secondary-ion mass spectrometry (TOF–SIMS) instruments, a 14-kDa poly(3,3,3-trifluoropropylmethylsiloxanes) fluorinated siloxane was selected for ion-NIMS.

In contrast with recent approaches focused on altering the composition of the ion beam by using ion clustering[15], adding matrix[16], or Ag cationization[17] to generate intact ions, NIMS has taken advantage of the inherent properties of nanostructured surfaces, for example the ability to trap liquids or polymers and to absorb light, and their low thermoconductivity. Laser irradiation of NIMS surfaces results in the generation of intact ions derived from a wide range of biomolecules such as proteins, peptides, metabolites and drugs (Fig. 1d). Unlike conventional MALDI, laser-NIMS is capable of generating multiply charged proteins similarly to electrospray ionization (ESI)[8] and cryo-infrared MALDI[18] currently with best results for proteins with a molecular mass of less than 30 kDa (Supplementary Fig. 4). Comparison of NIMS and ESI spectra of identical samples (Fig. 1d) reveals that the charge distribution of the protein is shifted to higher charge states for ESI, indicating that the protein is not as denatured by NIMS or possibly loses some charge during the ionization process. The inherent properties of nanostructured surfaces are also reflected in the higher sensitivity than with ESI and MALDI (Fig. 1d and Supplementary Figs 2, 3 and 5). Figure 1d shows the detection of 700 ymol of verapamil (a calcium antagonist drug) and a peptide mass fingerprint of 500 amol of bovine serum albumin, resulting in 55% sequence coverage. Taking advantage of the sensitivity of laser-NIMS it was possible to desorb/ionize endogenous phospholipids from few cells or even a single cell (Fig. 2 and Supplementary Discussion) of a highly metastatic breast cancer cell line (MDA-MB-231) with superior complexity than was achieved with hundreds of cells by nanoESI, MALDI or ion-NIMS.

The high lateral resolution (about 150 nm), reduced fragmentation, and sensitivity of ion-NIMS allow the direct characterization of peptide microarrays. *In situ* mass spectrometry enables label-free characterization of arrays and may eventually allow the analysis of biomolecules bound to the arrays from complex mixtures such as serum. Given the small feature size of current microarray

technologies (10 μm)[19], high-lateral-resolution analysis with SIMS is desirable; however, the fragmentation complicates identification[20]. Using ion-NIMS for high-resolution label-free analysis of a peptide array resulted in mass spectra and mass images obtained at 1 fmol of peptide (Fig. 3a, b), which is a 1,000-fold enhancement in sensitivity over other TOF–SIMS strategies for intact biomolecules[16,21] and doubles the mass range for peptide analysis with cluster sources[15]. This improvement in molecular ion sensitivity is probably due to a decrease in fragmentation typical of TOF–SIMS seen in identical samples spotted on a silicon control surface (Fig. 3d). The high resolution of NIMS revealed that bradykinin peptides localize themselves at the centre of the printed spot in depositions of higher concentration (Fig. 3c). However, because ion-NIMS is a surface effect, it may not be suitable for depth profiling experiments.

Unlike existing mass spectrometry imaging methods, which desorb/ionize molecules from the tissue surface[9,22], the NIMS surface permits the imaging of metabolites at the tissue/surface interface as shown in Fig. 4a. The initial step is ablation of the surface tissue with high laser energies (about 0.4 J cm^{-2} per pulse) followed by direct interrogation of the tissue/surface interface with lower-energy laser irradiation (about 0.01 J cm^{-2} per pulse) for mass spectral imaging (Fig. 4a). NIMS analysis of mouse embryo tissue sections reveals complex and different ion profiles between developing tissue types; some of these ions have been identified as lipids by using post-source decay because of their labile headgroups, for example phosphatidylcholine and phosphatidylethanolamine (Fig. 4, Supplementary Fig. 6 and Supplementary Table 2), whereas TOF/TOF or Fourier-transform mass spectrometry instrumentation should further enable the identification of other classes of metabolite. The high intensity of these phospholipids is attributed to both the high concentration of these species in biological membranes and the efficient partitioning of these metabolites onto the fluorinated clathrates (for example, bis(heptadecafluoro-1,1,2,2-tetrahydrodecyl)tetramethyldisiloxane) and not as a result of greater sensitivity of NIMS for lipids
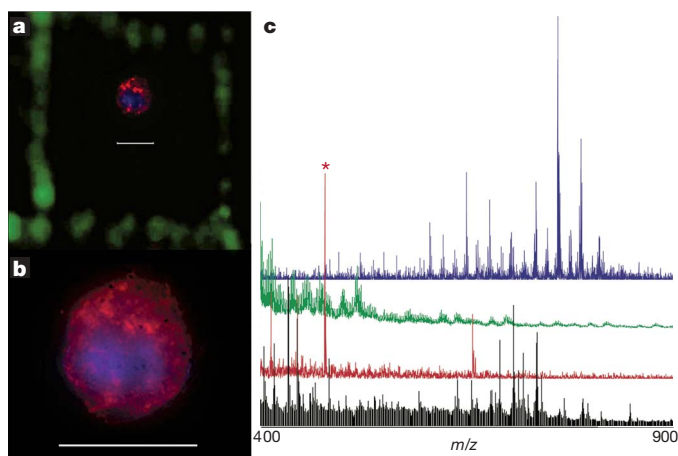

Figure 2 | Laser-NIMS and fluorescent analysis of a single cancer cell.
a, Fluorescent image of a MDA-MB-231 cancer cell after laser-NIMS analysis, fixing with paraformaldehyde, permeabilization with methanol, and staining with both 4,6-diamidino-2-phenylindole (blue) and Alexa 555-cytokeratin antibodies (red). The cell's location is indicated by being surrounded by a green fluorescent box formed with the same laser as that used for mass analysis at high intensity. **b**, Higher-resolution fluorescent image of the cell. **c**, Mass spectra from the cell with the use of laser-NIMS (blue) compared with 100 cells measured with ion-NIMS (green), MALDI analysis of 400 cells (red), and 500 cells measured with nanoelectrospray ionization (black). Laser-NIMS spectra were performed with bis(tridecafluoro-1,1,2,2-tetrahydrooctyl)tetramethyl-disiloxane initiator calibrated with Na$^+$, K$^+$, phosphocholine and 1-palmitoyllysophosphatidylcholine. Scale bar, 28 μm. Asterisk, matrix background ion.
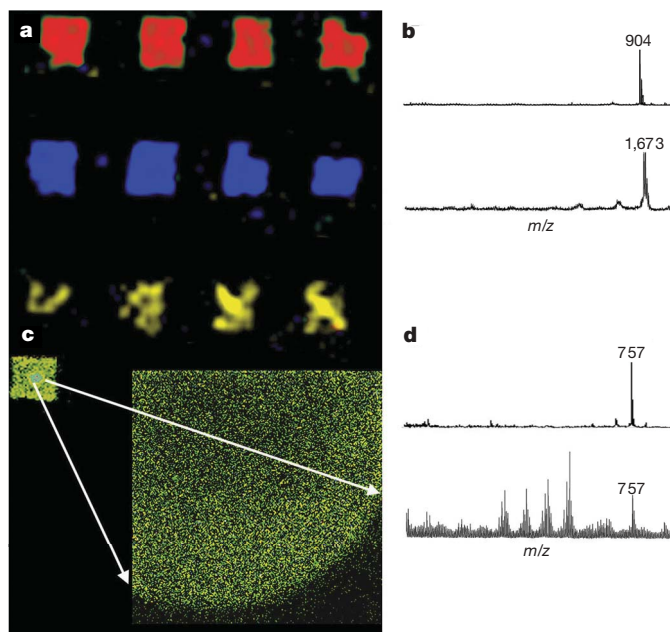

Figure 3 | Bi$_3^+$ ion imaging of peptide array on ion-NIMS surface.
a, Selective ion image of peptide array showing the localization of bradykinin 2-9 (red), neurotensin (blue) and bradykinin 1-7 (yellow) at 1 fmol. **b**, Corresponding mass spectra of bradykinin 2-9 (904 Da) and neurotensin (1,673 Da) at 1 fmol. **c**, Ion-NIMS mass spectrum of bradykinin 1-7 at 1 pmol revealing the localization of peptide in the centre of the spot as shown (inset) with a high-resolution image (150-nm ion beam). **d**, Bradykinin 1-7 (757 Da) also shown at 1 pmol both off the NIMS surface (top) and a silicon control surface, the latter showing significant fragmentation.
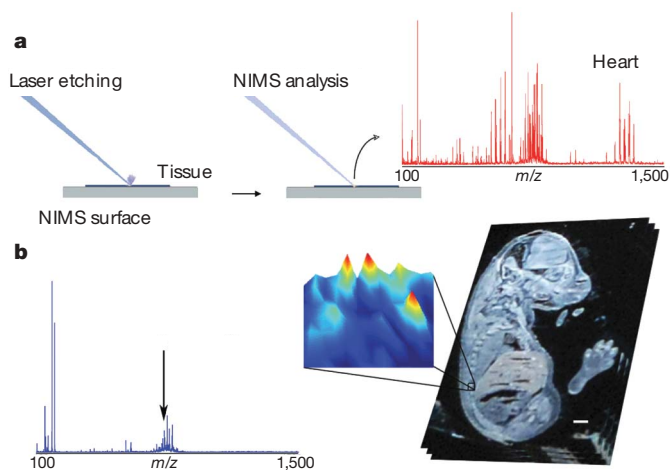
Figure 4 | Tissue imaging by laser-NIMS. a, The concept of tissue mass profiling: tissue slices (sections about 12 μm thick from a mouse embryo) on the NIMS surface are etched with high laser energies (about 0.4 J cm^{-2} per pulse), exposing the underlying surface. Laser irradiation (about 0.01 J cm^{-2} per pulse) of the 'etched' area results in desorption/ionization of metabolites, for example in developing heart tissue. **b,** Mass spectrometry imaging of glycerophosphatidylcholine (*m/z* 782, arrow) localized around a developing vertebra from a mouse embryo section (right). The resulting 300-μm-wide false-colour mass intensity image (inset) reveals a higher intensity (red, yellow and cyan) surrounding the developing vertebra (dark blue). The photograph is of the unstained section on the NIMS surface before analysis. The initiator was bis(heptadecafluoro-1,1,2,2-tetrahydrodecyl)tetramethyl-disiloxane and imaging was performed with a 20-μm step size by using a laser beam about 15 μm in diameter. Scale bar, 1 mm.
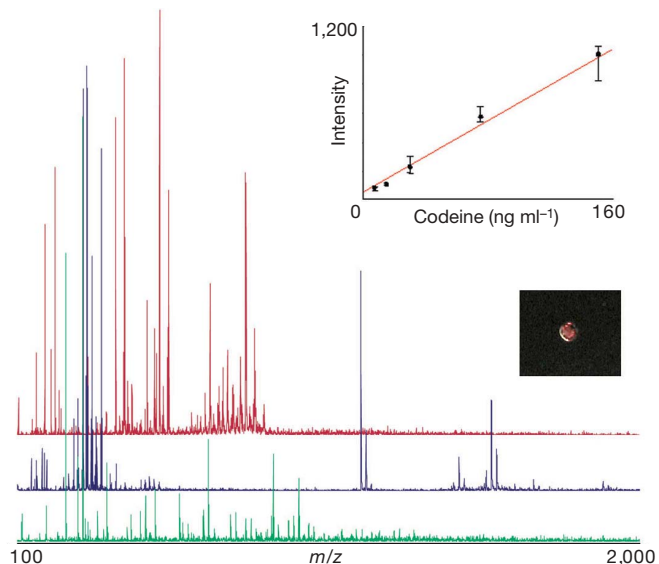

Figure 5 | Direct analysis of biofluids by laser-NIMS without sample preparation. Mass spectra of fresh blood (1 μl) obtained after laser etching (red) and urine (1 μl) (blue) with the use of bis(heptadecafluoro-1,1,2,2-tetrahydrodecyl)tetramethyl-disiloxane initiator. A markedly different spectrum (green) is obtained from the same urine by changing the initiator to 1,3-bis(hydroxybutyl)tetramethyldisiloxane. Inset: quantitative analysis of urine sample spiked with different concentrations of codeine shows a linear response up to 15 ng ml^{-1}, indicating that NIMS is suitable for metabolite quantification. Error bars represent s.d.

(Supplementary Fig. 7). In addition, changing to non-fluorous initiators changes the ion profiles markedly (Supplementary Fig. 8). Anatomical features are clearly visible, facilitating imaging of the developing vertebra, revealing the localization of the lipids around the developing bone (Fig. 4b and Supplementary Fig. 9), and the NIMS surface maintains metabolite localization on this length scale (about 50 μm). Overall, NIMS is well suited for imaging metabolites from tissues; in this it is complementary to MALDI, which is better suited to the detection of peptides and small proteins in tissue samples[23].

NIMS allows direct analysis (no sample preparation) of submicro-litre sample volumes of blood and urine, resulting in complex spectra without background ions (Fig. 5). This is in contrast with existing mass spectrometry approaches to biofluid analysis such as liquid chromatography (LC)–ESI, which requires extraction[24,25], and gas chromatography–mass spectrometry, which typically requires both extraction and derivatization[25]. Although MALDI can be used directly (a comparison of NIMS with MALDI is given in Supplementary Figs 10 and 11), fractionation and/or desalting are typically required to obtain good matrix–analyte co-crystallization[26,27]. Consistent with analysis of cells and tissues is the observation that lipids from blood effectively partition onto and ionize from perfluorinated bis(heptadecafluoro-1,1,2,2-tetrahydrodecyl)tetramethyl-disiloxane) clathrates. In contrast, urine spectra are largely dominated by small metabolites (less than 500 Da), although larger metabolites (such as putative gangliosides GM2 and GM1, with *m/z* 1545.9 and 1561.9, respectively) are also detected. Switching to a more polar hydroxy-lated siloxane initiator results in a markedly different ion profile, in which about 86% of the ions are unique to the perfluorinated surface. This suggests that initiator selection can be used to target metabolites on the basis of physical and chemical functionalities. In addition, because the concentration of metabolites is often critical in clinical assays, it is important to note that NIMS analysis is capable of detecting metabolites (in this case, codeine) quantitatively from urine at low levels (Fig. 5 inset; linear response until 15 ng ml^{-1}), in contrast with 75–250 ng ml^{-1} obtainable with the conventional approaches[28].

NIMS affords many possible applications and challenges: currently, our efforts are focused on interfacing NIMS with Fourier-transform mass spectrometry to improve metabolite identification on tissues, as well as immobilizing ligands on the NIMS surface for screening metabolite interactions and enzymatic activities. NIMS can also be developed for biomedical applications such as fundamental studies on single cells, developing high-throughput clinical tests, and conducting studies on pharmacokinetics and drug localization in tissues. The primary challenges include broadening the scope of NIMS through the use of new initiators, ideally including reactive initiators for direct chemical derivatization, and creating nanostructured surfaces to enhance desorption/ionization.

METHODS SUMMARY

Animal welfare. All procedures involving mice were approved by the Institutional Animal Care and Use Committee at Scripps Research Institute and conform to National Institutes of Health guidelines and public law.

Preparation of NIMS surfaces. Low-resistivity (0.01–0.02 Ω cm) P-type ⟨100⟩ (boron) silicon was etched at 48 mA cm^{-2} for 30 min. Various initiators were used on both silanized and unsilanized surfaces (Supplementary Information). For laser-NIMS, undiluted initiator solution was applied at room temperature (25 °C) for 30–60 min to the surface, and excess initiator was removed with a jet of nitrogen. For ion-NIMS, a 14-kDa poly(3,3,3-trifluoropropylmethylsiloxanes) polymer was applied to a silanized and vacuum-baked surface, then baked overnight at 100 °C; excess polymer was removed with a jet of nitrogen and brief rinsing with tert-butyl methyl ether.

Additional details are given in Supplementary Movies 2–8 and additional illustrated procedures in Supplementary Information.

Sample preparation. Peptides were printed by using robotic contact printing (1.4-nl drops). MDA-MB-231 cells were deposited on the NIMS surface with a micromanipulator, and after mass spectrometry analysis they were fixed and permeabilized, stained immunofluorescently and imaged. Embryos were collected, quickly frozen in OCT compound (VWR International), sectioned (12-μm slices) and then thaw-mounted on a NIMS chip. Fresh urine or blood samples were spotted directly on the NIMS surface and the excess was blown off with nitrogen. Urine samples were analysed directly and blood was first laser etched (see Fig. 5).

1. Want, E. J., Cravatt, B. F. & Siuzdak, G. The expanding role of mass spectrometry in metabolite profiling and characterization. *ChemBioChem* **6,** 1941–1951 (2005).
2. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422,** 198–207 (2003).
3. Chaurand, P., Stoeckli, M. & Caprioli, R. M. Direct profiling of proteins in biological tissue sections by MALDI mass spectrometry. *Anal. Chem.* **71,** 5263–5270 (1999).
4. Lindon, J. C., Holmes, E. & Nicholson, J. K. Metabonomics in pharmaceutical R&D. *FEBS J.* **274,** 1140–1151 (2007).
5. Chace, D. H. *et al.* Rapid diagnosis of phenylketonuria by quantitative analysis for phenylalanine and tyrosine in neonatal blood spots by tandem mass spectrometry. *Clin. Chem.* **39,** 66–71 (1993).
6. Karas, M. & Hillenkamp, F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal. Chem.* **60,** 2299–2301 (1988).
7. Tanaka, K. *et al.* Protein and polymer analyses up to *m/z* 100 000 by laser ionization time of flight mass spectrometry. *Rapid Commun. Mass Spectrom.* **2,** 151–153 (1988).
8. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246,** 64–71 (1989).
9. Caldwell, R. L. & Caprioli, R. M. Tissue profiling by mass spectrometry: a review of methodology and applications. *Mol. Cell. Proteomics* **4,** 394–401 (2005).
10. Benninghoven, A. in *Secondary Ion Mass Spectrometry* (eds Benninghoven, A., Rüdenauer, F. G. & Werner, H. W.) 1–6, 664–761 (Wiley, New York, 1987).
11. Kraft, M. L., Weber, P. K., Longo, M. L., Hutcheon, I. D. & Boxer, S. G. Phase separation of lipid membranes analyzed with high-resolution secondary ion mass spectrometry. *Science* **313,** 1948–1951 (2006).
12. Ostrowski, S. G. *et al.* Secondary ion MS imaging of lipids in picoliter vials with a buckminsterfullerene ion source. *Anal. Chem.* **77,** 6190–6196 (2005).
13. McLean, J. A., Stumpo, K. A. & Russell, D. H. Size-selected (2–10 nm) gold nanoparticles for matrix assisted laser desorption ionization of peptides. *J. Am. Chem. Soc.* **127,** 5304–5305 (2005).
14. Wei, J., Buriak, J. M. & Siuzdak, G. Desorption–ionization mass spectrometry on porous silicon. *Nature* **399,** 243–246 (1999).
15. Winograd, N. The magic of cluster SIMS. *Anal. Chem.* **77,** 142A–149A (2005).
16. Wu, K. J. & Odom, R. W. Matrix-enhanced secondary ion mass spectrometry: A method for molecular analysis of solid surfaces. *Anal. Chem.* **68,** 873–882 (1996).
17. Grade, H. & Cooks, R. G. Secondary ion mass-spectrometry—cationization of organic-molecules with metals. *J. Am. Chem. Soc.* **100,** 5615–5621 (1978).
18. Overberg, A., Karas, M., Bahr, U., Kaufmann, R. & Hillenkamp, F. Matrix-assisted infrared-laser (2.94-μm) desorption ionization mass-spectrometry of large biomolecules. *Rapid Commun. Mass Spectrom.* **4,** 293–296 (1990).
19. Singh-Gasson, S. *et al.* Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nature Biotechnol.* **17,** 974–978 (1999).
20. Hashimoto, H., Nakamura, K., Takase, H., Okamoto, T. & Yamamoto, N. Quantitative TOF-SIMS imaging of DNA microarrays produced by bubble jet printing technique and the role of TOF-SIMS in life science industry. *Appl. Surf. Sci.* **231–2,** 385–391 (2004).
21. Xu, J. Y., Braun, R. M. & Winograd, N. Applicability of imaging time-of flight secondary ion MS to the characterization of solid-phase synthesized combinatorial libraries. *Anal. Chem.* **75,** 6155–6162 (2003).
22. Wiseman, J. M., Ifa, D. R., Song, Q. & Cooks, R. G. Tissue imaging at atmospheric pressure using desorption electrospray ionization (DESI) mass spectrometry. *Angew. Chem. Int. Edn Engl.* **45,** 7188–7192 (2006).
23. Chaurand, P., Norris, J. L., Cornett, D. S., Mobley, J. A. & Caprioli, R. M. New developments in profiling and imaging of proteins from tissue sections by MALDI mass spectrometry. *J. Proteome Res.* **5,** 2889–2900 (2006).
24. Hao, C., Clement, R. & Yang, P. Liquid chromatography–tandem mass spectrometry of bioactive pharmaceutical compounds in the aquatic environment—a decade's activities. *Anal. Bioanal. Chem.* **387,** 1247–1257 (2007).
25. Maurer, H. H. Position of chromatographic techniques in screening for detection of drugs or poisons in clinical and forensic toxicology and/or doping control. *Clin. Chem. Lab. Med.* **42,** 1310–1324 (2004).
26. Fiedler, G. M. *et al.* Standardized peptidome profiling of human urine by magnetic bead separation and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Clin. Chem.* **53,** 421–428 (2007).
27. Villanueva, J. *et al.* Serum peptide profiling by magnetic particle-assisted, automated sample processing and MALDI-TOF mass spectrometry. *Anal. Chem.* **76,** 1560–1570 (2004).
28. Kraemer, T. & Paul, L. D. Bioanalytical procedures for determination of drugs of abuse in blood. *Anal. Bioanal. Chem.* **388,** 1415–1435 (2007).

METHODS

Materials. Hydrofluoric acid, ethanol, trifluoracetic acid, trichloroacetic acid, acetonitrile and methanol were purchased from Fisher Scientific and were of the highest purity available. All silanes and siloxanes were purchased from Gelest. Single-side-polished silicon wafers, P/boron-type, $\langle 100 \rangle$ orientation, low resistivity ($0.01–0.02 \,\Omega\, cm$), $525 \pm 25 \,\mu m$ thick and $100 \,mm$ in diameter, were obtained from Silicon Valley Microelectronics. Bovine serum albumin (BSA), thymopoietin II, sinapic acid, α-cyano-4-hydroxycinnamic acid, dithiothreitol, tert-butyl methyl ether, iodoacetamide, verapamil, β-lactoglobulin, 1-palmitoyllysophosphatidylcholine, neurotensin, bradykinin 1-7 and bradykinin 2-9 were from Sigma-Aldrich. Mass spectrometry grade porcine trypsin was from Promega. Bacteriorhodopsin was a gift from V.-P. Jaakola and R. C. Stevens. Fresh blood and urine from humans were used in this study. Amphetamine, methamphetamine, morphine, benzoylecgonine and codeine were gifts from R. L. Fitzgerald and D. Herold.

Instrumentation. An etching chamber was constructed in house and current was provided with a Bio-Rad PowerPack1000 (see details at http://masspec.scripps.edu/Research/nims/create.php). MDA-MB-231 cells (ATCC HTB-26) were grown at $37\,^{\circ}C$ in Leibovitz's L-15 medium with $2 \,mM$ L-glutamine and 10% calf bovine serum, then harvested from growth flasks by treatment with trypsin-EDTA. Human AB serum (Sigma) was used to block the surface prior to incubation with a mouse monoclonal anti-Pan cytokeratin antibody (Sigma) which was followed by incubation with Alexa555 goat anti-mouse antibody (Invitrogen) and 4,6-diamidino-2-phenylindole (DAPI) nuclear counterstain (Invitrogen).

Scanning electron microscopy images were collected on native (SiH) and NIMS surfaces coated with about 3 nm of Au/Pd with a Sirion FEG Digital Scanning Microscope (FEI Company). Laser-NIMS and MALDI were performed with a VoyagerDE STR (Applied Biosystems) equipped with delay extraction (DE) and a 337-nm pulsed nitrogen laser. Ion-NIMS was performed with ION-TOF IV TOF–SIMS (IONTOF GmbH) with a polyatomic Bi^+ source. An Agilent 1100 series LC–ESI-TOF was used for model compound analysis with direct infusion. BSA digest and cancer cell analysis were performed with a ThermoFinnigan LTQ linear ion trap fitted with a custom nanoelectrospray source (2 kV at the tip).

A micromanipulator (MINJ-PD; Tritech Research) was used to precisely count and expel cells onto the NIMS substrate, MALDI plate and Eppendorf tubes (nanoLC–ESI analysis). Cells were dried and mass spectral analysis was completed. Fluorescent imaging was performed with a Nikon TE2000 fluorescent microscope with an automated stage; a $20\times$ Plan Fluor (numerical aperture (NA) 0.45) objective was used for initial image acquisition and a $60\times$ oil Plan Apo (NA 1.40) objective for higher-resolution imaging. The microscope had automated excitation and emission filter wheels (Lamda 10-2; Sutter Instrument) with a triple band filter set for DAPI, fluorescein isothiocyanate and tetramethyl rhodamine isothiocyanate (61000V2; Chroma Technology Corp.). Digital images were acquired through a Retiga EXi Fast 1394 Mono Cooled digital camera (Qimaging).

Embryo tissue sections were prepared by cryosectioning with a Leica CM3050 S cryostat to a thickness of $12 \,\mu m$, and then thaw-mounted directly onto a NIMS chip.

Preparation of nanostructured silicon. Nanostructured silicon was prepared, with extreme caution, by using anodic etching in 25% hydrofluoric acid in ethanol. A current density of $48 \,mA \,cm^{-2}$ was applied for 30 min. After etching, undiluted initiator was applied at room temperature for 30–60 min on both silanized or unsilanized surfaces. Excess polymer was removed by using a jet of nitrogen (see more details at Supplementary Methods and http://masspec.scripps.edu/Research/nims/create.php for a step-by-step procedure). For laser-NIMS, different siloxane compounds were used as initiators (see Supplementary Table 1) and for ion-NIMS 14-kDa poly(3,3,3-trifluoropropylmethylsiloxanes) polymer was applied to the surface and baked overnight at $100\,^{\circ}C$; excess polymer was removed with a jet of nitrogen and brief dropwise rinsing with tert-butyl methyl ether.

Mass spectrometry. Laser-NIMS and MALDI: samples (for example, tissue slices) were revealed in the mass spectrometer by using a charge-coupled device camera. Each recorded mass spectrum resulted from the average of 100 laser shots on the area of interest. Acquisition parameters were set as follows: reflector mode, positive polarity, acceleration voltage 20 kV, first grid voltage 60%, guide wire voltage 0.15%, extraction delay time 100 ns. MALDI analysis was performed with a saturated solution of α-cyano-4-hydroxycinnamic acid in a water and acetonitrile (1:1) containing 0.1% trifluoracetic acid.

Ion NIMS: the pulse primary ion source was angled at 45° to the sample and was operated at 25 kV beam energy, with primary ion dosage below static limit, a 50-ns pulse width and calibrated off hydrocarbon peaks. Secondary ions were electrically directed to the time-of-flight tube by a sample stage at ± 2.5 kV and by an extraction lens oppositely biased at 4.5 kV.

NanoLC–ESI linear ion trap: nano-electrospray capillary columns (5 cm in length, $75 \,\mu m$ in inner diameter, containing 5-nm particles) tips were made in house with a P-100 laser puller (Sutter Instruments) packed with Zorbax SB-C_{18} stationary phase (Agilent). Water and acetonitrile (0.1% formic acid) were used as the mobile phases and the flow rate was $300 \,nl \,min^{-1}$. The sample was loaded directly onto the capillary. The gradient went from 5% acetonitrile to 30% acetonitrile in 10 min, then to 98% acetonitrile over 35 min, held at 98% acetonitrile for 10 min, and finally maintained for 10 min before re-equilibration at 5% acetonitrile.

Peptide array and single cells. Peptides were printed onto the NIMS surface from 1:1 acetonitrile/water solution containing 0.1% trifluoracetic acid by using a custom arrayer with an ArrayIt contact printing head. BSA was reduced and alkylated before digestion. Cells were placed onto the NIMS surface with a micromanipulator and left to dry before mass analysis. After mass spectrometry analysis, cells were fixed with 2% paraformaldehyde and permeabilized with cold methanol. Blocking of non-specific binding sites was performed with 20% human AB serum before cells were indirectly immunofluorescently stained with a mouse monoclonal anti-Pan cytokeratin antibody and Alexa555 goat anti-mouse antibody along with a DAPI nuclear counterstain.

Metabolite extraction from cells. Cell lysate was extracted with $600 \,\mu l$ of 86% methanol/13% water/1% formic acid. Samples were sonicated for 20 min and subsequently transferred to 2-ml Eppendorf tubes. The samples were placed in the freezer at $-20\,^{\circ}C$ and kept there for 60 min, after which they were centrifuged for 10 min in a 5417C centrifuge (Eppendorf) at 13,000 r.p.m.; the supernatant was decanted and dried before analysis.

Animals. C57Bl/6J mice were housed in ventilated cages in the Scripps Research Institute's vivarium, maintained on a 12-h dark/light cycle in a temperature- and humidity-controlled room. The care of the animals was in accord with the Lawrence Livermore National Laboratory (LLNL) Institutional Animal Care and Use Committee (IACUC) committee guidelines

Mouse embryo tissue slice preparation. Embryos were collected from pregnant dams at embryonic day 16.5, with the morning of vaginal plug detection designated embryonic day 0.5. Mice were deeply anaesthetized by isofluorane inhalation and killed by cervical dislocation. Embryos were collected, quickly frozen in OCT compound embedding medium (VWR International) on solid CO_2, equilibrated to $-20\,^{\circ}C$, sectioned at $12 \,\mu m$ and collected on the NIMS surface. Representative sections were collected on Superfrost Plus slides (Fisher Scientific) and stained with haematoxylin and eosin with routine histological techniques.

Blood and urine. Fresh urine and blood samples were spotted directly onto the NIMS surface without any further sample preparation. After 30 s, urine and blood drops were blown off with a nitrogen stream. For data processing of urine samples, all mass spectra were converted to ASCII text and imported into SpecAlign (http://physchem.ox.ac.uk/~jwong/specalign/) for baseline correction, normalization by total ion current and realignment/recalibration of individual spectra. For statistical analysis of codeine peak intensities, SigmaPlot 10 (Systat Software) was used.

Metabolite identification. Putative metabolites were identified on the basis of post-source decay fragment ions and the exact mass (m/z) by using the METLIN (http://metlin.scripps.edu/), Human Metabolome Project (http://redpoll.pharmacy.ualberta.ca/hmdb/HMDB/) and Lipid Maps (http://www.lipidmaps.org/) databases.

Non-equilibrium degassing and a primordial source for helium in ocean-island volcanism

Helge M. Gonnermann[1]† & Sujoy Mukhopadhyay[1]

Radioactive decay of uranium and thorium produces ^{4}He, whereas ^{3}He in the Earth's mantle is not produced by radioactive decay and was only incorporated during accretion—that is, it is primordial[1]. ^{3}He/^{4}He ratios in many ocean-island basalts (OIBs) that erupt at hotspot volcanoes, such as Hawaii and Iceland, can be up to sixfold higher than in mid-ocean ridge basalts (MORBs). This is inferred to be the result of outgassing by melt production at mid-ocean ridges in conjunction with radiogenic ingrowth of ^{4}He, which has led to a volatile-depleted upper mantle (MORB source) with low ^{3}He concentrations and low ^{3}He/^{4}He ratios[2–6]. Consequently, high ^{3}He/^{4}He ratios in OIBs are conventionally viewed as evidence for an undegassed, primitive mantle source, which is sampled by hot, buoyantly upwelling deep-mantle plumes[3,6,7]. However, this conventional model provides no viable explanation of why helium concentrations and elemental ratios of He/Ne and He/Ar in OIBs are an order of magnitude lower than in MORBs. This has been described as the 'helium concentration paradox'[8] and has contributed to a long-standing controversy about the structure and dynamics of the Earth's mantle. Here we show that the helium concentration paradox, as well as the full range of noble-gas concentrations observed in MORB and OIB glasses, can self-consistently be explained by disequilibrium open-system degassing of the erupting magma. We show that a higher CO_2 content in OIBs than in MORBs leads to more extensive degassing of helium in OIB magmas and that noble gases in OIB lavas can be derived from a largely undegassed primitive mantle source.

The high ^{3}He/^{4}He ratios in OIBs form the cornerstone of the conventional view that parts of the deep mantle have been isolated from outgassing and the convective upper mantle over Earth's history[5]. Preservation of an undegassed mantle reservoir is, however, difficult to reconcile with whole-mantle convection. In conjunction with the helium concentration paradox[8] this has led to alternative interpretations of high ^{3}He/^{4}He ratios in OIBs. For example, high ^{3}He/^{4}He ratios have been assigned to recycled lithosphere instead of primitive mantle[9] and it has been suggested that high ^{3}He/^{4}He ratios in OIBs indicate low ^{4}He rather than high ^{3}He concentrations[10]. Distinguishing between these and similar hypotheses is fundamental to our understanding of mantle dynamics and the origin of chemical heterogeneities in the mantle.

Because primitive mantle has considerably higher primordial noble-gas concentrations than the MORB mantle source[1], OIBs are expected to have higher ^{3}He concentrations than MORBs, but similar He/Ne and He/Ar elemental ratios[1,11]. However, it has been known for more than two decades that most OIB samples have significantly lower helium concentrations and also elemental ratios than MORBs[12]. This apparent discrepancy, the helium concentration paradox, cannot be explained by existing models of magma degassing[13,14] (Fig. 1). Constraining the ^{3}He concentration in parental (that

is, undegassed) OIB magma can provide a critical test for the presence of a primitive component in the OIB mantle source. Here we present a new, self-consistent model for degassing of CO_2, H_2O and noble gases (^{3}He, ^{4}He, ^{22}Ne, ^{21}Ne and ^{40}Ar) in OIB and MORB magmas, which resolves the helium concentration paradox and constrains the noble-gas concentrations in the parental magmas.

Volatile solubilities decrease during ascent-driven magma decompression, resulting in the growth of gaseous bubbles. If there is sufficient time for volatiles to diffuse from the melt into bubbles, the concentration of dissolved volatiles will be in equilibrium with the coexisting bubbles and the process is called equilibrium degassing. Degassing can be further classified as open-system or closed-system, depending on whether or not the exsolved bubbles are lost from the
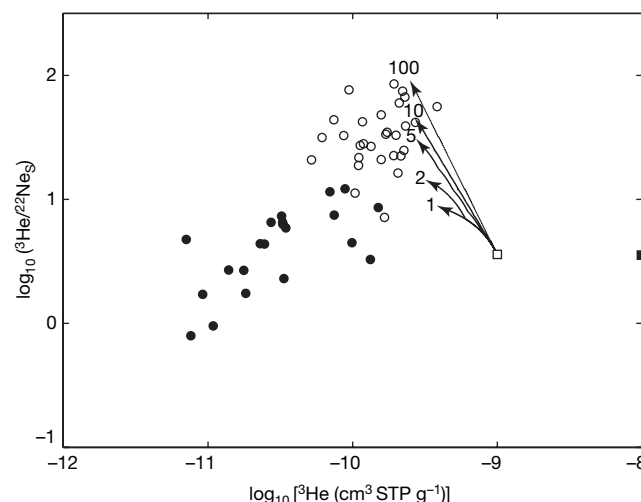

Figure 1 | Equilibrium degassing trajectories. Open and filled circles are MORB and OIB step heating analyses of submarine glasses, respectively (details are given in Supplementary Information). ^{22}Ne$_S$ is the mantle-derived primordial neon obtained by correcting the measured ^{22}Ne for atmospheric contamination[11]. Both the elemental ratio, ^{3}He/^{22}Ne$_S$, and ^{3}He concentrations in OIBs are considerably lower than in MORBs, contrary to expectations for OIBs derived from a primitive mantle source. Initial ^{3}He is 10^{-9} cm^3 STP g^{-1} (ref. 1). Curved arrows represent degassing paths for the melt. The shortest degassing path is for a closed-system equilibrium case (one degassing step). The remaining trajectories are for 2-step, 5-step, 10-step and 100-step open-system equilibrium degassing models. For details on degassing calculations see the text and Supplementary Information. ^{3}He concentration for undegassed OIB (filled square) is one order of magnitude higher than for MORB (open square). Starting from undegassed OIB will simply shift the degassing trajectories to the right and therefore none of the equilibrium degassing models can explain the low ^{3}He/^{22}Ne$_S$ or the low ^{3}He concentrations in OIBs.

[1]Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts 02138, USA. †Present address: Department of Geology and Geophysics, SOEST, University of Hawaii, Honolulu, Hawaii, 96822, USA.

magma during ascent and eruption. Most MORBs and OIBs are thought to erupt by means of open-system degassing[15,16]. During equilibrium degassing, elemental fractionation is controlled solely by the relative solubilities of the volatile species. Because noble-gas solubility decreases with increasing atomic size[17], He/Ar and He/Ne ratios always increase during equilibrium degassing and neither open-system nor closed-system equilibrium degassing can explain why erupted OIB lavas have elemental ratios lower than the starting composition[13,14] (Fig. 1). This is a fundamental problem, which any successful degassing model needs to explain, and is independent of whether the OIB mantle source has a higher volatile content than the MORB source.

Recent work has suggested that differences in diffusivity between noble gases may also lead to elemental fractionation in MORBs[16,18,19], in a process called disequilibrium degassing[20]. We propose that the difference in noble-gas concentrations and elemental ratios between MORBs and OIBs is a result of disequilibrium degassing. Disequilibrium degassing kinetically limits diffusive gas loss from the melt, and the dissolved concentrations of slowly diffusing volatiles may remain above equilibrium during degassing (Fig. 2). The
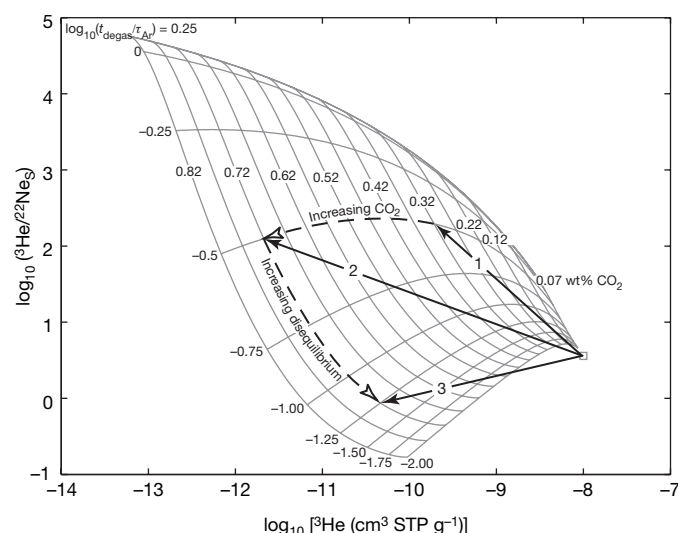

Figure 2 | Conceptual framework of our disequilibrium degassing model. Field of possible solutions for a degassing model with ten steps of open-system gas loss, an eruption depth of 5,500 m below sea level, and incorporating the effect of major volatiles on degassing at 55 MPa. Any point that falls within the solution grid represents the calculated dissolved noble-gas content of the degassed melt for a specific initial CO_2 and H_2O content and degassing time, but at a constant pressure. Because H_2O solubility is much higher than CO_2 solubility[21], most H_2O remains dissolved on eruption and does not significantly affect model results. We therefore always set initial H_2O content equal to initial CO_2. Degassing times are expressed in terms of the characteristic diffusion time of argon (the slowest diffusing noble gas under consideration) in basalt melt, τ_{Ar}. If t_{degas}, the time available for degassing during magma eruption, is similar to τ_{Ar}, the characteristic diffusion time of argon, then the degassing calculations are close to equilibrium. If $t_{degas} \ll \tau_{Ar}$, there is insufficient time for diffusion of argon (and also neon, because of its low diffusivity) and degassing is in disequilibrium. Because τ_{Ar} represents an e-folding time, concentrations are essentially at equilibrium for $\log_{10}(t_{degas}/\tau_{Ar}) > 0.25$. Subhorizontal lines are isopleths of constant degassing time relative to τ_{Ar}, with decreasing durations from top to bottom. Subvertical lines are isopleths of constant initial CO_2 (H_2O) with increasing concentrations from right to left. Solid arrows represent three illustrative degassing trajectories and mark the composition of the erupted lavas. Trajectories 1 and 2 have different initial CO_2 content but the same degree of disequilibrium. Trajectories 2 and 3 have the same initial CO_2 content but different degrees of disequilibrium. Note that higher initial CO_2 content leads to more helium loss from the melt (subhorizontal dashed arrow). Increasing the degree of disequilibrium, by decreasing the time available for degassing, results in lower He/Ne ratios (subvertical dashed arrow). Open square, undegassed melt.

magnitude of kinetic fractionation depends on magma vesicularity, bubble size and t_{degas}, the time available for degassing, which we express in terms of $\tau_{Ar} = a^2/D_{Ar}$, the characteristic diffusion time of argon in basalt melt. Here $D_{Ar} = 10^{-11.5}$ m^2 s^{-1} is argon diffusivity and $a = R(\phi^{-1/3} - 1)$ is the characteristic thickness of melt surrounding bubbles of radius R at a vesicle volume fraction ϕ. Kinetic fractionation of noble gases depends on relative diffusivities, which to first order scale as $D_{Ar} \sim D_{Ne} \sim 0.01 D_{He}$ (details are given in Supplementary Information). In other words, $\tau_{Ar} \sim \tau_{Ne} \sim 100 \tau_{He}$ and if $\tau_{Ar} > t_{degas} > 0.01 \tau_{Ar}$, then helium will be lost from the melt as pressure and, hence, solubility decrease. At the same time degassing of argon and neon will be limited by the ability of these gases to diffuse out of the melt. As a result, disequilibrium degassing will lead to He/Ne and He/Ar ratios in the erupted lavas that are lower than the starting composition (Fig. 2).

In our model we explicitly account for the effect of CO_2 and H_2O on degassing of noble gases. Varying the major volatile content of the magma affects the partial pressures of noble gases, and hence their solubility. Because most H_2O remains dissolved in the melt at typical eruptive pressures[21], we simplify our analysis and assume that H_2O concentrations in the undegassed magma are equal to CO_2 concentrations. When both the effects of disequilibrium and major volatiles are modelled quantitatively, helium concentrations in erupted lavas decrease as initial CO_2 increases, whereas neon and argon remain relatively unaffected, resulting in low He/Ne and He/Ar ratios (Fig. 2). We also account for eruption depth, because solubility depends on pressure; a lower eruption depth will increase the magnitude of helium degassing[22]. However, variations in eruption depth alone cannot resolve the helium concentration paradox (details are given in Supplementary Information).

We apply our model of magma degassing to MORBs and OIBs to investigate whether parental magmas of OIBs can have higher noble-gas abundances than MORBs, and whether the low elemental ratios in the erupted OIB lavas are the result of disequilibrium degassing. On eruption, volatile solubilities are lowest and magma ascent rates may be sufficient for disequilibrium to occur. Thus, to provide the simplest but still insightful analysis, we model only syneruptive magma degassing. To limit the number of free parameters and facilitate direct comparison between MORBs and OIBs, we model degassing of a single representative MORB and OIB magma. For all cases we assume syneruptive, open-system degassing with ten repeated steps of gas loss at the pressure of eruption onto the sea floor. Realistic variations of all parameters (for example, number of degassing steps, solubility constants, and diffusivities) affect the details of model predictions but not the overall conclusions of our work.

Our model parental MORB and OIB magmas have ^{4}He/^{21}Ne and ^{4}He/^{40}Ar ratios equal to time-integrated production ratios of 2.22×10^7 and 2.9, respectively[11], and ^{3}He/^{22}Ne ratios of 3.6 (refs 1, 23). Although MORB and OIB mantle sources may have different ^{3}He/^{22}Ne ratios[24] and ^{4}He/^{40}Ar may vary in the parental magmas[11], we choose the same starting composition to minimize free parameters. The ^{3}He concentration of the parental MORB magma is 10^{-9} cm^3 STP g^{-1} (cm^3 of gas at standard temperature and pressure per gram of melt) with a ^{3}He/^{4}He ratio of $8.1 R_A$ (measured ^{3}He/^{4}He relative to atmospheric)[11]. Our parental MORB magma has similar noble-gas concentrations to those of popping rock, 2ΠD43 (refs 25, 26), whose unusual volatile content may be representative of undegassed MORB[27], but could also be the consequence of bubble accumulation[25].

The volatile content of the OIB mantle source is thought to include contributions from both primitive and recycled material[28]. We choose a ^{3}He concentration for our parental OIB magma on the basis of generating a ^{3}He/^{4}He ratio of $24 R_A$, the average of our OIB data compilation, by mixing 80% recycled slab containing no ^{3}He and 20% primitive mantle with 4×10^{-9} cm^3 STP g^{-1} of ^{3}He (details are given in Supplementary Information). The resultant ^{3}He concentration of 10^{-8} cm^3 STP g^{-1} is an order of magnitude higher than in

parental MORB. OIBs with higher $^3He/^4He$ ratios require a smaller slab component in their mantle source, resulting in higher helium concentrations, and vice versa. Within realistic ranges of initial concentrations and elemental ratios our degassing model can match the entire OIB data array (Fig. 3).

Disequilibrium in MORBs has been observed previously[16,18,19] and the entire MORB data array can be explained by disequilibrium open-system degassing. For an eruption depth of 3,500 m below sea level, Fig. 3a, b illustrates how varying the CO_2 content of the parental MORB magma and the degassing time can capture the range of observed compositions in erupted lavas. Figure 4 shows our estimate of the parental CO_2 content for individual samples. Model calculations are for an eruption depth equal to the sample depth. The CO_2 content of the parental magma and t_{degas} were calculated by varying these parameters to minimize the least-squares error of predicted to measured ^3He, ^4He, ^{21}Ne, ^{22}Ne and ^{40}Ar concentrations (details are given in Supplementary Information). Concentration estimates for CO_2 of parental MORBs range from 0.07 to 0.24 wt% with a mean of 0.13 wt%, within the range of estimates for undegassed MORBs[29,30]. Average characteristic eruption times range from $0.05\tau_{Ar}$ to $0.23\tau_{Ar}$ with a mean of $0.12\tau_{Ar}$. Within a single eruption most bubbles will span a size range of 100 µm $\leq R \leq$ 1,000 µm (refs 19, 25). Reasonable values for vesicularity are $0.01 \leq \phi \leq 0.1$ (refs 19, 25). Thus, for

MORBs that ascend a distance of the order of 1,000 m, kinetic fractionation of helium from neon and argon will occur at ascent rates of 0.001 to 1 m s^{-1}, within the range of estimated values[16,31].

The entire OIB data array can also be explained by open-system degassing (Fig. 3c, d) with similar degrees of kinetic fractionation to those of MORBs. Figure 4 shows our estimates for CO_2 in parental OIB magmas, which range from 0.11 to 0.63 wt% with a mean of 0.36 wt%, also within the range of measured values[15]. Characteristic degassing times range from $0.09\tau_{Ar}$ to $0.20\tau_{Ar}$ with a mean of $0.15\tau_{Ar}$. Although shallow eruption depths of some OIBs ($\sim$1,000 m below sea level) give model fits with a similar initial CO_2 content to that of MORBs, some of the deepest OIB samples (>4,500 m below sea level) have very low helium concentrations and require high CO_2 contents (Fig. 4). Thus, OIBs require, on average, a higher initial volatile content than MORBs (Fig. 4) but no other significant differences in degassing.

$CO_2/^3He$ ratios from vent fluids are commonly assumed to reflect the composition of the parental magma[32,33]. However, during degassing the $CO_2/^3He$ ratios increase as helium is lost from the magma. $CO_2/^3He$ ratios of our parental magmas are about 10^9 for MORBs and 2×10^8 for OIBs, but degassing can produce more than one order of magnitude change in $CO_2/^3He$. Fractionation of CO_2 and ^3He is larger for OIBs, as a consequence of their higher CO_2 content and resultant increase in helium loss. Therefore, even in the absence
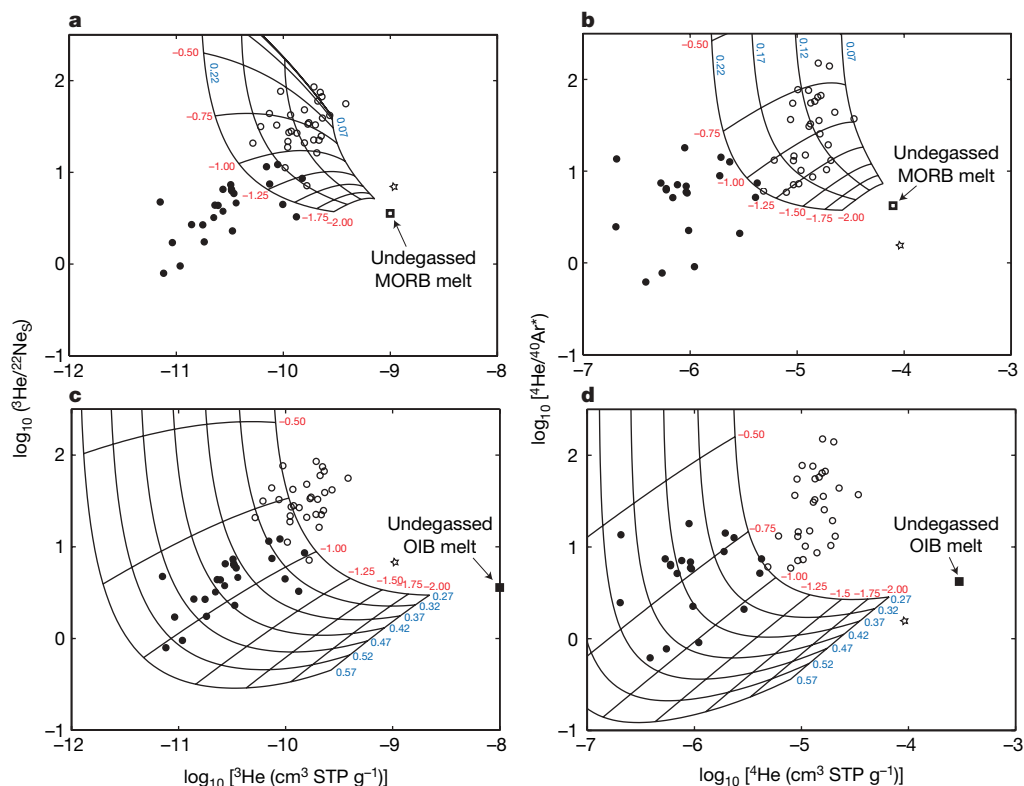

Figure 3 | Model results for disequilibrium degassing of MORBs and OIBs.
a, b, Solution grid for degassing of a MORB parental melt (open square); **c, d,** solution grid for degassing of an OIB parental melt (filled square). CO_2 numbers on the grid represent the initial CO_2 (H_2O) content in the undegassed melt (wt% CO_2; blue numbers). Negative isopleth labels (red numbers) represent the degree of disequilibrium and are values of $\log_{10}(t_{degas}/\tau_{Ar})$. Pressure is held constant and equivalent to an eruption depth of 3,500 m below sea level. Data are identical to those in Fig. 1. **a,** Plot of $^3He/^{22}Ne_S$ against ^3He in MORBs. Disequilibrium degassing models capture the range of compositions observed in erupted MORB lavas. **b,** Plot of $^4He/^{40}Ar^*$ against ^4He in MORBs. $^{40}Ar^*$ is the mantle-derived radiogenic argon obtained by correcting the measured ^{40}Ar for atmospheric contamination[11]. The range of solutions capture the MORB data array. Overall these model results suggest that MORBs are consistent with variable degrees of disequilibrium open-system degassing at characteristic eruption

times of the order of $0.1\tau_{Ar}$ from a parental MORB melt with low CO_2 and H_2O content. These results are also consistent for $^4He/^{21}Ne^*$ against ^4He (details are given in Supplementary Information). Open square, undegassed MORB melt. **c,** Plot of $^3He/^{22}Ne_S$ against ^3He in OIBs. **d,** Plot of $^4He/^{40}Ar^*$ against ^4He in OIBs. The solution grid encompasses all OIB data and produces lower noble-gas contents than erupted MORBs, as well as lower elemental ratios than the undegassed OIB melt (filled square). Solutions in He–Ar space are shifted slightly downwards relative to He–Ne space. A small decrease in $^3He/^{22}Ne_S$ of the parental OIB melt would compensate for this. Overall, our model results suggests that OIBs are consistent with disequilibrium open-system degassing with characteristic eruption times of the order of $0.1\tau_{Ar}$, and CO_2 (H_2O) contents in the parental magmas that are higher than for MORBs. These results are also consistent for $^4He/^{21}Ne^*$ against ^4He space (details are given in Supplementary Information). Open circles, MORBs; filled circles, OIBs; open star, 2ΠD43.

1039
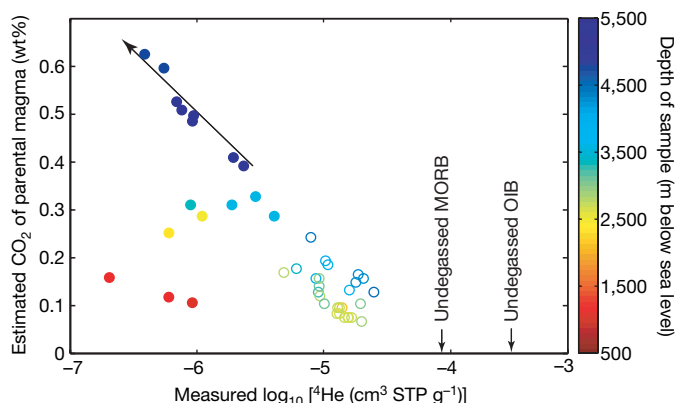
Figure 4 | Predicted CO₂ content of parental magmas. Plot of CO_2 against [4]He in MORBs (open circles) and OIBs (filled circles), with colours representing sample depth. Vertical arrows indicate the composition of the parental MORB and OIB magmas. Using the degassing model, we estimate the CO_2 content and degree of disequilibrium required to provide the minimum misfit (details are given in Supplementary Information) between predicted and measured noble-gas concentrations (^{3}He, ^{4}He, ^{22}Ne$_s$, ^{21}Ne* and ^{40}Ar*) for each individual sample at an eruption depth equivalent to the reported sample depth. The initial CO_2 content of MORBs ranges from 0.07 to 0.24 wt% with a mean of 0.13 wt%. The predicted CO_2 content of OIBs ranges from 0.11 to 0.63 wt% with a mean of 0.36 wt%. Overall, OIBs require a higher and wider range of initial CO_2, with concentrations up to about 0.7 wt%. Predicted eruption times (not shown) range from $0.05\tau_{Ar}$ to $0.23\tau_{Ar}$ with a mean of $0.12\tau_{Ar}$ for MORBs, and $0.09\tau_{Ar}$ to $0.20\tau_{Ar}$ with a mean of $0.15\tau_{Ar}$ for OIBs. The diagonal arrow indicates that increasing the CO_2 content of the undegassed magma decreases partial pressure and solubility of helium during degassing.

of fractionation during hydrothermal circulation[33], observed $CO_2/^3$He ratios in vent fluids and lavas may not be representative of the mantle source.

Thus, because of relatively high diffusivity, degassing of helium is not significantly limited by diffusion. Higher CO_2 content of OIBs dilutes helium in the exsolved gas phase, relative to MORBs, thereby lowering partial pressures and helium solubility. This explains why OIB lavas have lower helium abundances than MORBs, even though the OIB parental magma is more helium rich. Degassing of neon and argon is limited by diffusion in both MORBs and OIBs. In combination with higher helium loss in OIBs, this explains the low He/Ne and He/Ar ratios of OIBs relative to MORBs (Fig. 3). For example, disequilibrium open-system degassing of parental OIB magmas with noble-gas abundances of 1×10^{-8}, 2.8×10^{-9} and 7.1×10^{-5} cm^3 STP g^{-1} for ^{3}He, ^{22}Ne and ^{40}Ar, respectively, is consistent with observations (Figs 3 and 4).

We conclude that there is no helium concentration paradox. Low noble-gas concentrations and elemental ratios of OIBs are simply the consequence of disequilibrium degassing of a more volatile-rich parental magma. Our results are consistent with high ^{3}He/^{4}He ratios in OIB lavas being derived from a primitive mantle component with ^{3}He, ^{22}Ne and ^{36}Ar abundances of the order of 10^{-9} cm^3 STP g^{-1}. In agreement with some previous estimates[28], the percentage of primitive mantle in the OIB source is about 20% or less, but our calculations do not constrain its size, geometry or location. Regardless of whole-mantle convection, it seems that part of the Earth's mantle has remained largely undegassed over Earth's history.

1. Porcelli, D. & Ballentine, C. J. Models for the distribution of terrestrial noble gases and evolution of the atmosphere. *Rev. Mineral. Geochem.* **47**, 411–480 (2002).
2. Hart, R., Dymond, J. & Hogan, L. Preferential formation of the atmosphere–sialic crust system from the upper mantle. *Nature* **278**, 156–159 (1979).
3. Kurz, M. D., Jenkins, W. J. & Hart, S. R. Helium isotopic systematics of oceanic islands and mantle heterogeneity. *Nature* **297**, 43–47 (1982).
4. Allègre, C. J., Staudacher, T., Sarda, P. & Kurz, M. Constraints on evolution of Earth's mantle from rare-gas systematics. *Nature* **303**, 762–766 (1983).

5. O'Nions, R. K. & Oxburgh, E. R. Heat and helium in the Earth. *Nature* **306**, 429–431 (1983).
6. Kellogg, L. H. & Wasserburg, G. J. The role of plumes in mantle helium fluxes. *Earth Planet. Sci. Lett.* **99**, 276–289 (1990).
7. Morgan, W. J. Convective plumes in lower mantle. *Nature* **230**, 42–43 (1971).
8. Anderson, D. L. The helium paradoxes. *Proc. Natl Acad. Sci. USA* **95**, 4822–4827 (1998).
9. Helfrfrich, G. R. & Wood, B. J. The Earth's mantle. *Nature* **412**, 501–507 (2001).
10. Anderson, D. L. A model to explain the various paradoxes associated with mantle noble gas geochemistry. *Proc. Natl Acad. Sci. USA* **95**, 9087–9092 (1998).
11. Graham, D. W. Noble gas isotope geochemistry of mid-ocean ridge and ocean island basalts: Characterization of mantle source reservoirs. *Rev. Mineral. Geochem.* **47**, 247–317 (2002).
12. Fisher, D. E. Noble gases from oceanic island basalts do not require an undepleted mantle source. *Nature* **316**, 716–718 (1985).
13. Honda, M. & Patterson, D. B. Systematic elemental fractionation of mantle-derived helium, neon, and argon in mid-oceanic ridge glasses. *Geochim. Cosmochim. Acta* **63**, 2863–2874 (1999).
14. Moreira, M. & Sarda, P. Noble gas constraints on degassing processes. *Earth Planet. Sci. Lett.* **176**, 375–386 (2000).
15. Dixon, J. E. & Clague, D. A. Volatiles in basaltic glasses from Loihi seamount, Hawaii: Evidence for a relatively dry plume component. *J. Petrol.* **42**, 627–654 (2001).
16. Paonita, A. & Martelli, M. A new view of the He–Ar–CO₂ degassing at mid-ocean ridges: Homogeneous composition of magmas from the upper mantle. *Geochim. Cosmochim. Acta* **71**, 1747–1763 (2007).
17. Carroll, M. R. & Stolper, E. M. Noble gas solubilities in silicate melts and glasses: New experimental results for argon and the relationship between solubility and ionic porosity. *Geochim. Cosmochim. Acta* **57**, 5039–5051 (1993).
18. Burnard, P., Harrison, D., Turner, G. & Nesbitt, R. Degassing and contamination of noble gases in Mid-Atlantic Ridge basalts. *Geochem. Geophys. Geosyst.* **4**, 10.1029/2002GC000326 (2003).
19. Aubaud, C., Pineau, F., Jambon, A. & Javoy, M. Kinetic disequilibrium of C, He, Ar and carbon isotopes during degassing of mid-ocean ridge basalts. *Earth Planet. Sci. Lett.* **222**, 391–406 (2004).
20. Lensky, N. G., Navon, O. & Lyakhovsky, V. Bubble growth during decompression of magma: experimental and theoretical investigation. *J. Volcanol. Geotherm. Res.* **129**, 7–22 (2004).
21. Dixon, J. E. Degassing of alkalic basalts. *Am. Mineral.* **82**, 368–378 (1997).
22. Hilton, D. R. et al. Controls on magmatic degassing along the Reykjanes Ridge with implications for the helium paradox. *Earth Planet. Sci. Lett.* **183**, 43–50 (2000).
23. Honda, M. et al. Possible solar noble-gas component in Hawaiian basalts. *Nature* **349**, 149–151 (1991).
24. Honda, M. & McDougall, I. Primordial helium and neon in the Earth—a speculation on early degassing. *Geophys. Res. Lett.* **25**, 1951–1954 (1998).
25. Javoy, M. & Pineau, F. The volatiles record of a 'popping' rock from the Mid-Atlantic Ridge at 14° N: chemical and isotopic composition of gas trapped in the vesicles. *Earth Planet. Sci. Lett.* **107**, 598–611 (1991).
26. Moreira, M., Kunz, J. & Allègre, C. J. Rare gas systematics in popping rock: Isotopic and elemental compositions in the upper mantle. *Science* **279**, 1178–1181 (1998).
27. Sarda, P. & Graham, D. Mid-ocean ridge popping rocks: implications for degassing at ridge crests. *Earth Planet. Sci. Lett.* **97**, 268–289 (1990).
28. Ballentine, C. J., van Keken, P. E., Porcelli, D. & Hauri, E. H. Numerical models, geochemistry and the zero-paradox noble-gas mantle. *Phil. Trans. R. Soc. Lond. A* **360**, 2611–2631 (2002).
29. Holloway, J. R. Graphite-melt equilibria during mantle melting: constraints on CO₂ in MORB magmas and the carbon content of the mantle. *Chem. Geol.* **147**, 89–97 (1998).
30. Saal, A. E., Hauri, E. H., Langmuir, C. H. & Perfit, M. R. Vapour undersaturation in primitive mid-ocean-ridge basalt and the volatile content of Earth's upper mantle. *Nature* **419**, 451–455 (2002).
31. Head, J. W., Wilson, L. & Smith, D. K. Mid-ocean ridge eruptive vent morphology and substructure: Evidence for dike widths, eruption rates, and evolution of eruptions and axial volcanic ridges. *J. Geophys. Res.* **101**, 28265–28280 (1996).
32. Hilton, D. R., McMurtry, G. M. & Goff, F. Large variations in vent fluid CO₂/³He ratios signal rapid changes in magma chemistry at Loihi seamount, Hawaii. *Nature* **396**, 359–362 (1998).
33. Resing, J. A., Lupton, J. E., Feely, R. A. & Lilley, M. D. CO₂ and ³He in hydrothermal plumes: implications for mid-ocean ridge CO₂ flux. *Earth Planet. Sci. Lett.* **226**, 449–464 (2004).

Group formation stabilizes predator–prey dynamics

John M. Fryxell[1], Anna Mosser[2], Anthony R. E. Sinclair[3] & Craig Packer[2]

Theoretical ecology is largely founded on the principle of mass action, in which uncoordinated populations of predators and prey move in a random and well-mixed fashion across a featureless landscape. The conceptual core of this body of theory is the functional response, predicting the rate of prey consumption by individual predators as a function of predator and/or prey densities[1–5]. This assumption is seriously violated in many ecosystems in which predators and/or prey form social groups. Here we develop a new set of group-dependent functional responses to consider the ecological implications of sociality and apply the model to the Serengeti ecosystem. All of the prey species typically captured by Serengeti lions (*Panthera leo*) are gregarious, exhibiting nonlinear relationships between prey-group density and population density. The observed patterns of group formation profoundly reduce food intake rates below the levels expected under random mixing, having as strong an impact on intake rates as the seasonal migratory behaviour of the herbivores. A dynamical system model parameterized for the Serengeti ecosystem (using wildebeest (*Connochaetes taurinus*) as a well-studied example) shows that grouping strongly stabilizes interactions between lions and wildebeest. Our results suggest that social groups rather than individuals are the basic building blocks around which predator–prey interactions should be modelled and that group formation may provide the underlying stability of many ecosystems.

Although the adult population of Serengeti lions has shown considerable variation in total abundance over the past 27 years (ref. 6), most of this variation is accommodated by changes in the number of prides in a given area (Fig. 1; $y = -1.0939 + 0.173x$, $R^2 = 0.524$, $F_{1,26} = 28.59$, $P < 0.001$). As a consequence, the mean number of adult lions per pride (G) has not varied significantly with population abundance ($F_{1,26} = 0.002$, $P = 0.844$), averaging 6.3 adults per group. During periods when lions are abundant, variation in pride size tends to increase as a result of the successful recruitment of offspring as well as the formation of smaller prides by fission[6].

All eight of the lions' most common prey species are gregarious; thus, the rate at which lions encounter their prey depends on the density of herds rather than individuals. We estimated the lions' prey encounter rates by driving 33 ground censuses across the lion study area between 2004 and 2007, recording the density of prey groups per unit area and relating group density to individual density, with prey groups operationally defined as any individuals occurring within a 2-ha area. Although the frequency of encounter with prey groups was positively associated with prey density (Fig. 1), the best fit ($R^2 = 0.76–0.98$) was always provided by a simple power function ($y = cN^b$), where y is herd density per km², c and b are the intercept and exponent of the power curve, and N is the prey density per km². In each species, the best-fit estimate of the exponent b was significantly less than 1 (see Supplementary Information), implying that each doubling of prey density leads to a much more modest increase in encounter frequency with groups.

We used the Serengeti data to develop four alternative functional responses by lions feeding on wildebeest (Box 1): a null response, assuming that lions forage solitarily and prey are randomly distributed (equation (1) in Box 1), a grouped lion functional response (equation (2)), a grouped prey response (equation (3)), and a functional response assuming group formation both by lions and their prey (equation (4)). Group formation by either species depresses the attack rate at a given population density, but group formation by both lions and wildebeest is especially effective (Fig. 2).

If we assume that both species were solitary, local stability analysis confirms that a lion–wildebeest system with parameters taken from Serengeti would be unstable for most combinations of conversion efficiency (ε, which is defined as the rate with which wildebeest biomass is converted to lion biomass) and lion mortality rate (d) (Fig. 3a). The potential for local stability is increasingly enhanced by group formation by lions (Fig. 3b), by wildebeest (Fig. 3c) and by both lions and wildebeest (Fig. 3d). Nonlinear relationships between group density and prey population density are not essential for stabilization. For example, a linear relationship between wildebeest group density and population density data would be even more strongly stabilizing than the nonlinear best-fit model shown in Fig. 1. The key requirement for a stabilizing effect is that prey-group density is lower than individual density.
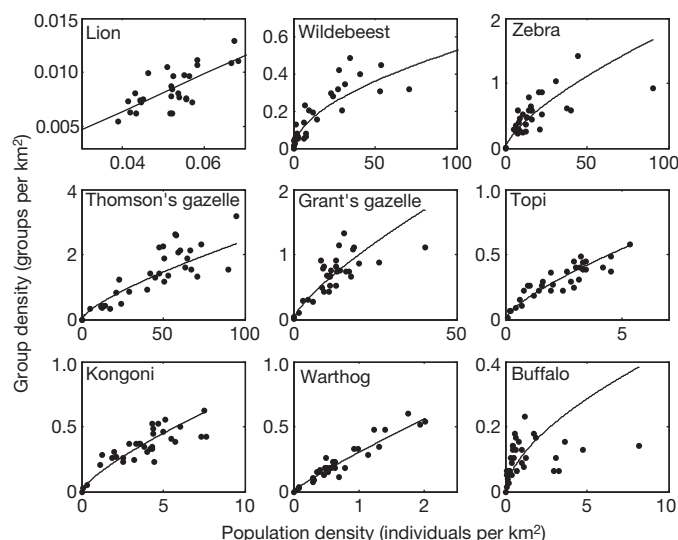

Figure 1 │ Group density in relation to population density for Serengeti lions and their predominant eight prey species. Herbivore prey include wildebeest (*Connochaetes taurinus*), zebra (*Equus burchelli*), Thomson's gazelle (*Gazella thomsoni*), Grant's gazelle (*Gazella granti*), topi (*Damaliscus korrigum*), kongoni (*Alcelaphus buselaphus*), warthog (*Phacochoerus aethiopicus*) and Cape buffalo (*Synerus caffer*). Lines show best-fit power functions for herbivores and the best-fit linear function for lions, based on least-squares minimization.

[1]Department of Integrative Biology, University of Guelph, 50 Stone Road East, Guelph, Ontario, Canada, N1G 2W1. [2]Department of Ecology, Evolution, and Behavior, University of Minnesota, 1987 Upper Buford Circle, St Paul, Minnesota 55108, USA. [3]Zoology Department, University of British Columbia, 6270 University Boulevard, Vancouver, British Columbia, Canada, V6T 1Z4.

Box 1 | Group-dependent functional responses

If lions forage solitarily, the expected type II functional response[1] is

$$\Psi(N) = \frac{aN}{1 + a(h_1 + h_2)N} \qquad (1)$$

where a is the area of effective search per unit time, h_1 is the expected time to attack and subdue each prey item, h_2 is the expected time to consume and digest each prey item, N is prey density per km^2, and $\Psi(N)$ is prey intake per predator per day.

Group formation by predators changes the handling time of prey (see Methods), leading to the functional response

$$\Psi(N,G) = \frac{aN}{G + a(Gh_1 + h_2)N} \qquad (2)$$

where G is the predator group size. We can evaluate the effect of prey group formation by inserting a modified encounter rate (acN^b) into the standard type II functional response (equation (1)), where $c = \exp(\text{intercept})$ and b is the slope of the linear regression of ln(prey group density) versus ln(prey density, N). This alters the functional response to

$$\Psi(N) = \frac{acN^b}{1 + a(h_1 + h_2)cN^b} \qquad (3)$$

Equations (2) and (3) can be combined to calculate the functional response when both predators and their prey form groups:

$$\Psi(N,G) = \frac{acN^b}{G + a(Gh_1 + h_2)cN^b} \qquad (4)$$

Herd formation reduces search efficiency by predators by creating 'holes' across the landscape that would otherwise be occupied by asocial prey, analogous to the effects of weak diffusive movement by predators and/or prey[7–14]. Sociality accordingly has important dynamical implications because of the reduced frequency with which predators encounter prey[3,13]. In contrast, group formation by predators limits search efficiency to a level similar to a solitary predator because of overlapping perceptual ranges. Although cooperation may compensate, at least to some degree, for the reduced capture rates implicit in the group-dependent functional response, the available evidence does not convincingly demonstrate such an effect[15]. Most individual lions refrain from contributing to group hunts except when pursuing Cape buffalo (*Synserus caffer*), which are inaccessible to solitary individuals or small groups[16]. The fact that natural selection has favoured sociality in lions despite the sizeable cost in terms of prey capture suggests that the compensating benefits must be profound. Lion sociality derives from the greater reproductive success of mothers that defend their cubs cooperatively against infanticidal males[17] and the advantages of group territoriality against neighbouring prides[18,19].

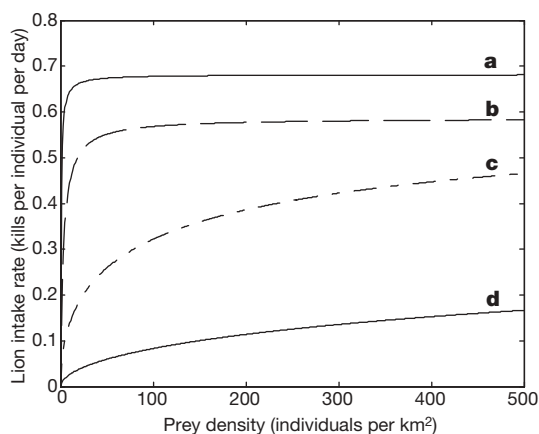**Figure 2 | Predicted effects of group formation by Serengeti lions and/or their prey on the functional response of individual lions. a**, Both predators and prey are solitary; **b**, only predators are gregarious; **c**, only prey are gregarious; **d**, both predators and prey are gregarious.

Previous Serengeti models have shown that herbivore migration provides a seasonal escape from lion predation[20]. Several important prey species (wildebeest, zebra (*Equus burchelli*), Thomson's gazelles (*Gazella thomsoni*) and Grant's gazelles (*Gazella granti*)) spend most of the year in areas far out of reach of a given lion pride. As a result of migratory movements, wildebeest densities varied considerably across the monthly censuses (see Supplementary Information). We assessed the impact of wildebeest movement on lion foraging by comparing average prey intake during censuses that were above the median group density of 0.205 wildebeest groups per km^2 and those that were below the median. In periods when wildebeest were abundant, lion feeding rates averaged 0.074 ± 0.038 wildebeest per day (mean $\pm$ s.d.; $n = 15$), whereas when wildebeest were scarce, lion feeding rates averaged 0.013 ± 0.0092 wildebeest per day ($n = 15$), a reduction of 82%.

Using wildebeest densities at carrying capacity ($N = 49.6\,\text{km}^{-2}$) as a point of reference, the group-dependent functional response (equation (4)) yields a kill rate of 0.064 wildebeest per lion per day, whereas the non-group-dependent functional response (equation (1)) predicts a kill rate of 0.674 wildebeest per lion per day, a reduction of 90%. The reduction in kill rates as a result of group formation (90%) is therefore similar to the effect of migration (82%). Group formation and migration are complementary processes, reducing the rate of wildebeest predation by lions by nearly two orders of magnitude when acting in tandem.

Previous demographic analyses of the Serengeti wildebeest suggested that starvation inflicted a greater magnitude of mortality than predation and that food limitation was density-dependent[21]. The time-series data show little indication of appreciable covariation in wildebeest and lion densities (Fig. 4), despite the fact that wildebeest contribute strongly to the diet of Serengeti lions; the lion data also do not show autocorrelation functions with significant lags, the fundamental statistical signature of cyclic population dynamics[22]. These results are most consistent with the hypothesis that Serengeti wildebeest are indeed regulated primarily by food abundance rather than by predation, and that the interaction between lions and wildebeest is relatively stable.

The observed patterns suggest that group formation and seasonal migration by wildebeest are potent behavioural mechanisms that contribute substantially to the stability of lion–wildebeest interactions in Serengeti. Lion densities remained constant for decade-long intervals owing to a relatively constant number of prides in each habitat; they then suddenly jumped to new 'equilibria' after major changes in prey abundance or availability (and an associated change
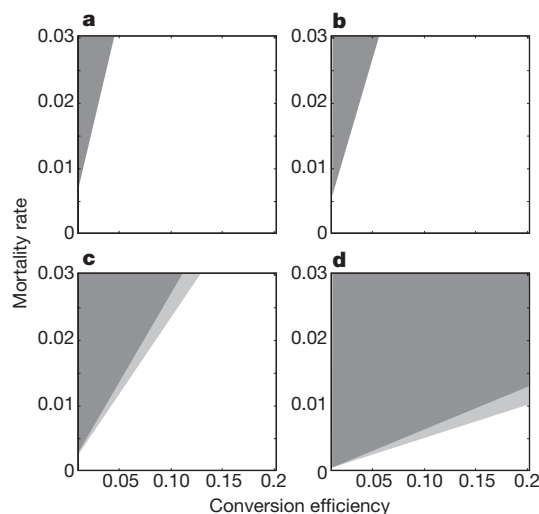

Figure 3 | Locally stable parameter combinations (ε versus d) for lion–wildebeest models. a, Both predators and prey are solitary; **b**, only predators are gregarious; **c**, only prey are gregarious; **d**, both predators and prey are gregarious. Locally stable combinations are lightly shaded, and locally unstable combinations are unshaded; heavily shaded combinations cannot sustain predators.
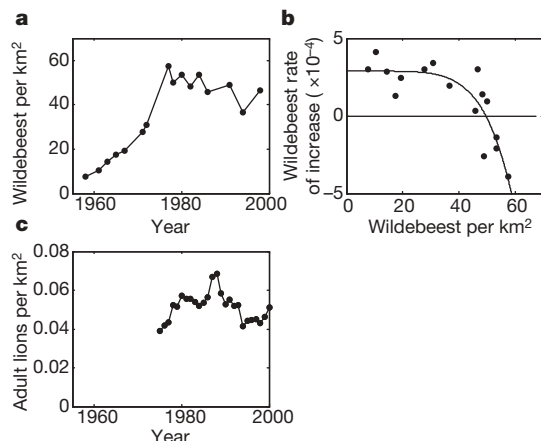
**Figure 4 | Time-series data for Serengeti wildebeest and lions.
a,** Population densities of wildebeest over time; **b,** the theta-logistic model of daily per capita rate of increase by wildebeest; **c,** population densities of lions over time. Time-series analysis (Akaike's Information Criteria scores calculated with the Levinson–Durbin algorithm for a Yule–Walker autoregressive process) showed no significant evidence of periodic variation in the abundance of either adult lions or the total lion population (order 1 favoured in each case).

in pride density)[6]. Our models clearly predict long periods of population stability for sedentary social predators that primarily feed on mobile herds of prey. Rather than stability being the exception, as predicted by classical theoretical studies of predator–prey interactions, stability may characterize most ecological communities owing to the almost universal tendency for prey species to form flocks, herds, swarms and schools and for so many predators to form packs, pods, prides and clans.

METHODS SUMMARY

Our general approach involved four linked steps. We first derived an a priori set of feeding rates (functional responses), based on four different assumptions: first, both predators and prey are solitary; second, predators are gregarious, whereas prey are solitary; third, prey are gregarious, whereas predators are solitary; and last, both predators and prey are gregarious. These functional response models were derived as variants of the type II functional response[1] that is most commonly applied in predator–prey models.

Parameters for the new functional response models were estimated for the Serengeti ecosystem by using field data for lions and the predominant eight species of herbivores preyed on by lions. These data include estimates of the amount of meat consumed by each lion, the time expenditure per hunt, the number of hunts required for each kill, the time expenditure per kill, the time expenditure for consumption of prey, the effective search radius for hunting lions, and the digestive pause in lions[16,23,24]. We also estimated the relationship between group density (groups per km[2]) and individual density (individuals per km[2]) for herbivores, based on 33 vehicle transects, each of 391 km, driven across the lion study area[6]. Herbivore transects were replicated at roughly monthly intervals during 2004–2007 to assess seasonal variation in prey abundance.

The alternative functional responses were incorporated into interactive predator–prey models[25,26]. These models were based on a theta-logistic model for the Serengeti wildebeest population[27] that best approximates the curvilinear density-dependent response (Fig. 4). We then used local stability analysis[28,29] to compare the range of stable versus unstable parameter combinations for each functional response form to evaluate the ecological impact of group formation on the population stability of predators and their prey.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 9 July; accepted 17 August 2007.

1. Holling, C. S. The components of predation as revealed by a study of small-mammal predation of the European pine sawfly. *Can. Entomol.* **91**, 293–320 (1959).

2. Arditi, R. & Ginzburg, L. R. Coupling in predator–prey dynamics: ratio-dependence. *J. Theor. Biol.* **139**, 311–326 (1989).
3. Cosner, C., DeAngelis, D. L., Ault, J. S. & Olson, D. B. Effects of spatial grouping on the functional response of predators. *Theor. Popul. Biol.* **56**, 65–75 (1999).
4. Jeschke, J. M., Kopp, M. & Tollrian, R. Predator functional responses: discriminating between handling and digesting prey. *Ecol. Monogr.* **72**, 95–112 (2002).
5. Abrams, P. A. & Ginzburg, L. R. The nature of predation: prey dependent, ratio dependent, or neither? *Trends Ecol. Evol.* **15**, 337–341 (2000).
6. Packer, C. *et al.* Ecological change, group territoriality, and population dynamics in Serengeti lions. *Science* **307**, 390–393 (2005).
7. McCauley, E., Wilson, W. & de Roos, A. M. Dynamics of age-structured and spatially-structured predator–prey interactions: individual-based models and population-level formulations. *Am. Nat.* **142**, 412–442 (1993).
8. Nisbet, R. M., de Roos, A. M., Wilson, W. G. & Snyder, R. E. Discrete consumers, small scale resource heterogeneity, and population stability. *Ecol. Lett.* **1**, 34–37 (1998).
9. Donalson, D. D. & Nisbet, R. M. Population dynamics and spatial scale: effects of system size on population persistence. *Ecology* **80**, 2492–2507 (1999).
10. Keeling, M. J., Wilson, H. B. & Pacala, S. W. Reinterpreting space, time lags, and functional responses in ecological models. *Science* **290**, 1758–1761 (2000).
11. Pascual, M., Mazzega, P. & Levin, S. A. Oscillatory dynamics and spatial scale: the role of noise and unresolved pattern. *Ecology* **82**, 2357–2369 (2001).
12. Hosseini, P. R. How localized consumption stabilizes predator–prey systems with finite frequency of mixing. *Am. Nat.* **161**, 567–585 (2003).
13. Nachman, G. A functional response model of a predator population foraging in a patchy habitat. *J. Anim. Ecol.* **75**, 948–958 (2006).
14. Cantrell, R. S. & Cosner, C. The effect of spatial heterogeneity on population dynamics. *J. Math. Biol.* **29**, 315–338 (2001).
15. Packer, C. & Ruttan, L. M. The evolution of cooperative hunting. *Am. Nat.* **132**, 159–198 (1988).
16. Scheel, D. & Packer, C. Group hunting behaviour of lions: a search for cooperation. *Anim. Behav.* **41**, 697–709 (1991).
17. Packer, C., Scheel, D. & Pusey, A. E. Why lions form groups: food is not enough. *Am. Nat.* **136**, 1–19 (1990).
18. McComb, K. E., Packer, C. & Pusey, A. E. Roaring and numerical assessment in contests between groups of female lions *Panthera leo. Anim. Behav.* **47**, 379–387 (1994).
19. Grinnell, J., Packer, C. & Pusey, A. E. Cooperation in male lions: kinship, reciprocity or mutualism? *Anim. Behav.* **49**, 95–105 (1995).
20. Fryxell, J. M., Greever, J. & Sinclair, A. R. E. Why are migratory ungulates so abundant? *Am. Nat.* **131**, 781–798 (1988).
21. Mduma, S. A. R., Sinclair, A. R. E. & Hilborn, R. Food regulates the Serengeti wildebeest population: a 40-year record. *J. Anim. Ecol.* **68**, 1101–1122 (1999).
22. Turchin, P. *Complex Population Dynamics* (Princeton Univ. Press, Princeton, NJ, 2003).
23. Scheel, D. Profitability, encounter rates, and prey choice of African lions. *Behav. Ecol.* **4**, 90–97 (1993).
24. Elliott, J. P., Cowan, I., McT. & Holling, C. S. Prey capture by the African lion. *Can. J. Zool.* **55**, 1811–1828 (1977).
25. Rosenzweig, M. L. Paradox of enrichment—destabilization of exploitation ecosystems in ecological time. *Science* **171**, 385–387 (1971).
26. Rosenzweig, M. L. & MacArthur, R. H. Graphical representation and stability conditions of predator–prey interactions. *Am. Nat.* **97**, 209–223 (1963).
27. Sinclair, A. R. E., Fryxell, J. M. & Caughley, G. *Wildlife Ecology, Conservation, and Management* (Blackwell, Oxford, 2006).
28. Yodzis, P. *Introduction to Theoretical Ecology* (Harper & Row, New York, 1989).
29. Hastings, A. *Population Biology: Concepts and Models* (Springer, New York, 1997).

Author Contributions Field data used for this paper came from long-term collaborative studies by J.M.F., C.P. and A.R.E.S., with statistical analysis of the data by J.M.F. and A.M. Models were developed and analysed by J.M.F. All four authors participated in discussion of the results at a workshop and contributed to manuscript preparation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to J.M.F. (jfryxell@uoguelph.ca).

METHODS

Data sources. Three sources of data were used to parameterize the group-dependent functional response by lions. First, Scheel[16,23] followed individual Serengeti lions during 96-h watches just before or after full moons, recording all hunting activity as well as counts of prey species in the immediate vicinity of focal individuals. These data provide estimates of the amount of meat consumed by each lion, the time expenditure per hunt, the number of hunts required for each kill, the time expenditure per kill, the time expenditure for consumption of prey, and the effective search radius for hunting lions. We augmented Scheel's data with the field estimate by Elliott *et al.*[24] of the digestive pause in lions, a key component of handling time, to predict the type II functional response (Box 1).

The second source of data was regular monitoring of lion prides in the 2,000-km² Serengeti study area since 1966; mean pride size was calculated annually across 27 years of data. Most hunting is conducted by adult females, and beginning in 1984 at least one female in each pride was fitted with a radio collar and relocated every few days. The lions' belly sizes were routinely recorded, and the overnight displacement distance was calculated for animals whose bellies were initially empty (because these individuals were the most motivated to search for prey; $n = 2,218$ occasions). The maximum overnight displacement ($10\,\mathrm{km\,d^{-1}}$) provided an estimate of velocity during prey search. The search path radius of 200 m was based on Scheel's statistical analysis[23] of the determinants of successful versus unsuccessful attacks.

Our third data source was encounter rates with herds in relation to herbivore density. To estimate this relationship, we conducted 33 monthly herbivore censuses during 2004–2007 over the same study area monitored by the lion project. Herbivores were counted from a vehicle travelling at 10–$20\,\mathrm{km\,h^{-1}}$ along each 391-km path, recording the frequency of encounters as well as herd size and species composition within a 100-m radius of the vehicle.

Functional response derivation. Here we consider the effects of group formation by lions and their prey on the rate of prey consumption. For simplicity and notational convenience, we consider only wildebeest, but the general pattern applies to all prey species. As described by Cosner *et al.*[3], a tight group of predators should experience an overlapping visual range; the area of effective search per unit time (a) of a group-hunting predator should therefore be equivalent to that of a solitary predator. We estimated the effective search area for any group size of lions by multiplying the effective search path ($\alpha = 0.4\,\mathrm{km}$) by the travel velocity ($\omega = 10\,\mathrm{km\,d^{-1}}$) and the probability of success per attack ($\sigma = 0.29$ for wildebeest). Hence, $a = \alpha\sigma\omega = 1.16\,\mathrm{km^2\,d^{-1}}$. The average time required by lions to capture each prey item (h_1) is a critical component of the total handling time and is estimated by multiplying the average time per attack by the expected number of attacks per successful capture[23]. For wildebeest, $h_1 = 0.045$ days, and Scheel[23] found no evidence that h_1 varied with lion group size. The expected digestive pause (h_2) is a direct linear function of meat per carcass. The estimated wildebeest carcass mass ($85.2\,\mathrm{kg}$)[23] combined with the digestive time estimates by Elliott *et al.*[24] (0.401 h per kilogram consumed) suggest that $h_2 = 1.422$ days for each wildebeest killed. The digestive time for each predator depends on predator group size, however, because each individual feeds from the same

carcass[3]. Hence, the digestive pause for a group of size G equals h_2/G. Finally, the capture rate per individual in the group is obtained by dividing the group capture rate by group size.

There are some circumstances in which a type II functional model may not provide an accurate model of consumption rate, even for solitary predators and prey, because of spatial heterogeneities in capture risks for individual prey[30]. Nonetheless, the type II response is commonly used as a 'null' model that captures the essential elements of many ecological interactions. Asymmetries in risk[30] would further reduce the efficiency of predation, although by far less than the order of magnitude caused by group formation.

Trophic model. We used a modified Rosenzweig–MacArthur model as our template for considering the effect of group formation on lion–wildebeest interactions[25,26], a model for which the dynamics has been thoroughly studied[28,29]. The model has the following structure for the rates of change by predators and prey:

$$\mathrm{d}N/\mathrm{d}t = r_{\max}N[1 - (N/K)^{\theta}] - \Psi(N,G)P \qquad (5)$$

$$\mathrm{d}P/\mathrm{d}t = \Psi(N,G)\varepsilon P - dP \qquad (6)$$

The modelled wildebeest population of density N has theta-logistic growth, where $r_{\max}$ is the maximum per capita rate of change of wildebeest, θ is a dimensionless parameter influencing the curvature of the relationship between the exponential rate of growth and wildebeest population density, and K is the equilibrium population density of wildebeest in the absence of predation. The lion population of density P grows at a per capita rate determined by d (the lion mortality rate), ε (a coefficient converting consumed wildebeest into new lion recruits) and $\Psi(N,G)$ (the lion functional response, using equations (1)–(4) derived in Box 1). Long-term population data were available (Fig. 4a) to arrive at a first approximation of the parameters for wildebeest logistic growth[27], suggesting that $r_{\max} = 2.884 \times 10^{-4}$ per day, $\theta = 5.946$ and $K = 49.6$ wildebeest per km² (based on the assumption that the more than one million wildebeest occupy the full Serengeti ecosystem, comprising 25,000 km²). The fit of this model is shown in Fig. 4b.

For all models of this form, stability properties depend on the magnitude of the elements of the community interaction matrix (α_{ij}), evaluated at equilibria for the system (N_{eq} and P_{eq}). This is due largely to changes in predator and prey equilibria, which change in a complex manner with predator and prey group size (see Supplementary Information). The predator–prey system will be locally stable if the eigenvalues of the community matrix have a negative real part[28,29]. This will apply only when the combination of community matrix coefficients are such that $\alpha_{11} + \alpha_{22} < 0$ and $\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21} > 0$, evaluated at the equilibria N_{eq} and P_{eq} (refs 28, 29). Expressions for the community matrix coefficients are given in Supplementary Information.

30. Mols, C. M. M. *et al.* Central assumptions of predator–prey models fail in a semi-natural experimental system. *Proc. R. Soc. Lond. B* **271** (Supplement), S85–S87 (2004).

LETTERS

Revisiting Lévy flight search patterns of wandering albatrosses, bumblebees and deer

Andrew M. Edwards[1]†, Richard A. Phillips[1], Nicholas W. Watkins[1], Mervyn P. Freeman[1], Eugene J. Murphy[1], Vsevolod Afanasyev[1], Sergey V. Buldyrev[2,3], M. G. E. da Luz[4], E. P. Raposo[5], H. Eugene Stanley[2] & Gandhimohan M. Viswanathan[6]

The study of animal foraging behaviour is of practical ecological importance[1], and exemplifies the wider scientific problem of optimizing search strategies[2]. Lévy flights are random walks, the step lengths of which come from probability distributions with heavy power-law tails[3,4], such that clusters of short steps are connected by rare long steps. Lévy flights display fractal properties, have no typical scale, and occur in physical[3–5] and chemical[6] systems. An attempt to demonstrate their existence in a natural biological system presented evidence that wandering albatrosses perform Lévy flights when searching for prey on the ocean surface[7]. This well known finding[2,4,8,9] was followed by similar inferences about the search strategies of deer[10] and bumblebees[10]. These pioneering studies have triggered much theoretical work in physics (for example, refs 11, 12), as well as empirical ecological analyses regarding reindeer[13], microzooplankton[14], grey seals[15], spider monkeys[16] and fishing boats[17]. Here we analyse a new, high-resolution data set of wandering albatross flights, and find no evidence for Lévy flight behaviour. Instead we find that flight times are gamma distributed, with an exponential decay for the longest flights. We re-analyse the original albatross data[7] using additional information, and conclude that the extremely long flights, essential for demonstrating Lévy flight behaviour, were spurious. Furthermore, we propose a widely applicable method to test for power-law distributions using likelihood[18] and Akaike weights[19,20]. We apply this to the four original deer and bumblebee data sets[10], finding that none exhibits evidence of Lévy flights, and that the original graphical approach[10] is insufficient. Such a graphical approach has been adopted to conclude Lévy flight movement for other organisms[13–17], and to propose Lévy flight analysis as a potential real-time ecosystem monitoring tool[17]. Our results question the strength of the empirical evidence for biological Lévy flights.

In 1992, five wandering albatrosses (*Diomedea exulans*) on Bird Island, South Georgia (54° 00′ S, 38° 03′ W), each had a salt-water immersion logger[21] attached to one of its legs. Over the course of a bird's foraging trip, the logger recorded the proportion of each hour spent sitting on the sea surface. In ref. 7, flight durations (time intervals between landing on the ocean) were then calculated as consecutive hours for which a bird remained dry, to a resolution of 1 h. It was assumed that birds landed on the water solely to feed, and that flight durations were thus indicative of distances between prey.

Time series for 19 separate foraging trips[7] were pooled to give a total of 363 flights. The resulting log–log histogram of flight durations gave a straight line with a slope of approximately 2, and is reproduced in Supplementary Fig. 1 from the original raw data. The crux of the conclusion that the albatrosses were performing Lévy flights was that the slope of 2 implied the probability density function of flight durations t (in hours) was[7,10]

$$f(t) \sim t^{-2} \qquad (1)$$

for $t \geq 1$ h (leaving out the normalization constant). This is consistent with the Lévy flight definition that the tail of the probability density function is of the power-law form $t^{-\mu}$, where $1 < \mu \leq 3$ (although technically this is a Lévy walk[4,7,22]). The Lévy flight was inferred to be an efficient foraging strategy for food that might be fractally distributed on the ocean surface[7].

We first analyse a newer, larger and higher resolution data set of albatross flight durations to test for Lévy flights. In 2004, 20 wandering albatrosses on Bird Island were each fitted with a salt-water logger and a GPS device. The GPS data were too infrequent (at most one location h^{-1}) to give distances between landings, but were needed to estimate each bird's departure time from Bird Island, in order to calculate the duration of the initial flight before first landing on the water (we calculated return flights similarly). The resulting data set of flight records was pooled, as in ref. 7, yielding a total of 1,416 flights to a resolution of 10 s (Fig. 1).

The flights ≥ 1 h are clearly inconsistent with coming from the power law t^{-2} ascertained[7] for the 1992 data. Furthermore, data from a power law of any exponent (not just 2) would yield a straight line[23], and this is clearly not the case. In fact, the flight durations t (in h) are consistent with coming from the shifted gamma distribution given by the probability density function

$$f(t) = \frac{r^s}{\Gamma(s)} y^{s-1} e^{-ry} \qquad (2)$$

where $y = t - 1/120$ accounts for the assumed 30-s period before the bird searches for new food sources (see Methods), $s = 0.31$ is the shape parameter, $r = 0.41$ h^{-1} is the rate parameter, and Γ is the gamma function. Equation (2) is valid for flights > 30 s; for shorter flights we have $f(t) = 0$. The exponential term of equation (2) dominates for large t, implying Poisson behaviour, such that for long enough flights the birds essentially encounter prey randomly with a constant low probability.

A brownian random walker's displacement increases as t^H where $H = 1/2$. If $H > 1/2$, we have 'superdiffusion', as originally inferred in Fig. 2a of ref. 7. Superdiffusion is explained by one or both of the following[3,4]: (1) the probability density function of flight lengths has a

[1]British Antarctic Survey, High Cross, Madingley Road, Cambridge CB3 0ET, UK. [2]Center for Polymer Studies and Department of Physics, Boston University, Boston, Massachusetts 02215, USA. [3]Department of Physics, Yeshiva University, 500 West 185th Street, Room 1112, Belfer Hall, New York, New York 10033, USA. [4]Departamento de Física, Universidade Federal do Paraná, C. P. 19044, 81531-990, Curitiba-PR, Brazil. [5]Laboratório de Física Teórica e Computacional, Departamento de Física, Universidade Federal de Pernambuco, 50670-901, Recife-PE, Brazil. [6]Instituto de Física, Universidade Federal de Alagoas, 57072-970, Maceió-AL, Brazil. †Present address: Pacific Biological Station, Fisheries and Oceans Canada, 3190 Hammond Bay Road, Nanaimo, British Columbia V9T 6N7, Canada.

heavy tail, or (2) there is long-range persistence in direction or time. The term Lévy flight is usually reserved[4] for case (1). Reference 7 used a Lévy walk model that assumes constant velocity, yielding a power law probability density function of flight durations. The gamma distribution (equation (2)) has $\mu = 1 - s = 0.69$. This is such a slow power-law decay that it is non-normalizable unless there is truncation. Hence, unlike the truncated Lévy flight[3] where $1 < \mu \leq 3$, equation (2) cannot be interpreted as a power law with exponential truncation. Superdiffusion remains possible, but through directional persistence only, not Lévy flights.

The longest flight in 2004 was 14.9 h, whereas for the original 1992 data[7], 25 of the 363 flights were >15 h. In 1992, for six of the trips the birds were also fitted with a satellite transmitter (Platform Terminal Transmitter, PTT), which provided locations at irregular intervals. Although these data are also too coarse to determine flight distances between landings (and were not available for ref. 7), we use them here to determine when each bird left and returned to Bird Island (see Methods). Figure 2 shows the wet/dry data for these six trips, together with the estimated departure and return times on the basis of the PTT data. For trip 3B, for example, the first dry sequence, based solely on the salt-water-logger data, is 46 h. However, the PTT data reveal that the bird did not leave Bird Island until 41 h after the logger was switched on. Thus, the true duration of the first flight was only 5 h.

For the remaining five trips, the original dry sequences from the loggers (in the order of Fig. 2) of 44 h, 69 h, 26 h, 67 h and 23 h represent, in reality, flight records of only 4 h, 3 h, 1 h, <1 h and <1 h, respectively. Similarly, for final flights the raw logger data values of 4 h, 8 h, 13 h, 9 h, 34 h and 9 h get corrected to true flight records of 4 h, 5 h, 8 h, 2 h, 3 h and 7 h, respectively.

However, in ref. 7 the raw logger data were assumed to represent true flights. We adjust the data for the remaining 13 trips, for which no PTT data were collected, by eliminating the initial and final dry sequences (see Methods). Using these adjusted data, in Fig. 3a we compute a corrected version of the original log–log histogram shown in Fig. 3a of ref. 7. There are now no flight durations in the two largest bins, and the longest flight is only 20 h compared to the original 99 h. The data thus no longer span two orders of magnitude, and the points
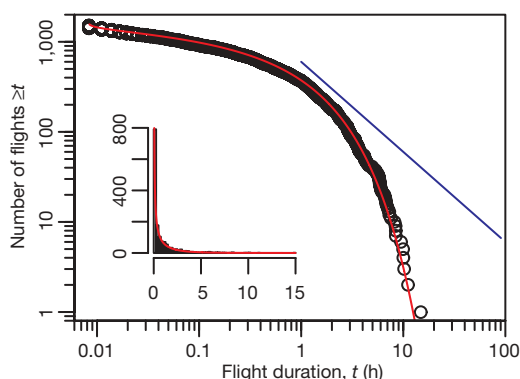
clearly lie on a curve, not a straight line that would indicate a power law.

The data are also consistent with coming from a shifted gamma distribution of the form of equation (2), illustrated in Fig. 3b. The resulting confidence intervals for s and r are much larger than those for the 2004 data, because of the smaller sample size, lack of data <1 h, and lower resolution of the data. Figure 3c shows the uncorrected and corrected data as a rank/frequency plot, as in Fig. 1. We conclude that, when time spent by the birds on the nests is accounted for, the original 1992 albatross data do not support Lévy flight behaviour.

The original albatross study[7] was followed by reports of Lévy flight behaviour by deer (*Dama dama*)[10] and bumblebees (*Bombus terricola*)[10]. The deer data were plotted as a standard histogram log-transformed (LT in the terminology of ref. 24), the bumblebee data as a smoothed histogram log-transformed[10], and straight lines were compared to the tails. The problematic[24,25] LT method was then used to conclude Lévy flight behaviour in other studies[13,15–17,26], and the logarithmic binning with normalization[24] (LBN) method used in ref. 14. In no cases were alternative distributions properly considered, or goodness-of-fit tests performed. In Box 1 we present a new method to overcome these shortcomings. We now illustrate this approach by re-analysing the deer and bumblebee data[10].

Following ref. 10, we digitized data from ref. 27 concerning foraging times of deer in unfenced and fenced areas, and digitized data from ref. 28 concerning flights of bumblebees between flower heads in high- and low-food situations. In ref. 10 these data were all assumed to relate to distances travelled between food items. Likelihood functions modified from those in Box 1 are calculated in the Supplementary Information (requiring numerical maximization) to account for the data only being available already binned. The resulting Akaike weights w_1 for the unbounded power-law tail were $<10^{-8}$ for both deer data sets, and 0.40 and 0.001 for the bees, respectively (for the 0.40 case, μ is outside the Lévy range). Given such negligible support for the unbounded power law in the Lévy range, we also tested both models over the bounded ranges $[a, b]$ assumed in ref. 10; see Table 1.

We find that for the deer scenarios the exponential distribution is favoured by Akaike weights of 0.9994 and 0.95. Furthermore, the
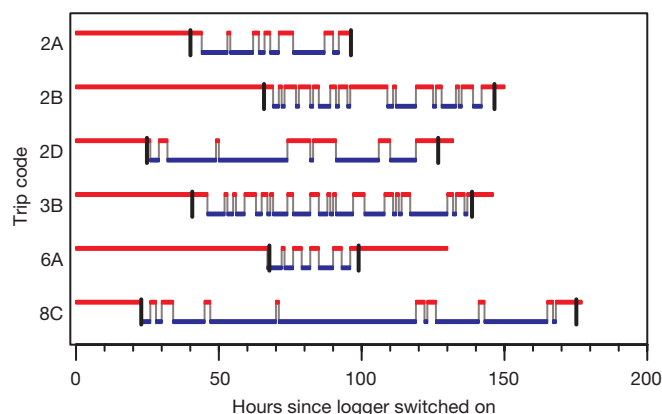

Figure 1 | Rank/frequency plot[23] of 2004 wandering albatross data, showing no evidence for Lévy flight behaviour. Circles show number of flights $\geq t$ for each flight duration t (calculated by ranking flights by size). The red curve is the fit to the shifted gamma distribution (equation (2)) with maximum likelihood estimates (MLEs) of $s = 0.31$ (95% confidence interval (CI): 0.27–0.34) and $r = 0.41$ h^{-1} (95% CI: 0.36–0.46), obtained by maximizing the multinomial likelihood function that takes into account the discrete sampling nature of the loggers (see Supplementary Information). The data are consistent with coming from this distribution ($n = 1,416$, degrees of freedom = 37, $G = 28.9$, $P = 0.83$). Flights are correct to within ± 10 s (see Supplementary Information). If the flights ≥ 1 h followed the power law with exponent $\mu = 2$ as in ref. 7, the points would lie on the straight blue line[23] (that has been vertically shifted slightly for clarity)—this is clearly not the case. The inset shows the 2004 data as a conventional histogram on linear axes, with number of flights against flight duration in hours.

Figure 2 | Data for the six wandering albatross trips in 1992 that have known departure and return times. Red lines are hours for which a logger was completely dry; blue lines indicate hours when a logger was wet for some part of that hour; and grey lines indicate switches between these regimes. Black lines indicate when each bird departed from and returned to Bird Island, on the basis of the PTT data. Time 0 is when the loggers were switched on at a computer—thus intervals before the first black lines include time taken to affix the logger to a bird plus time spent by the bird sitting on its nest before departing. All birds remained on Bird Island for long periods before departing, but such periods were considered to be flights in the original study[7]. Intervals after the final black lines correspond to time the bird sat on its nest after returning plus time spent retrieving the logger, but these were also originally considered as flights[7].
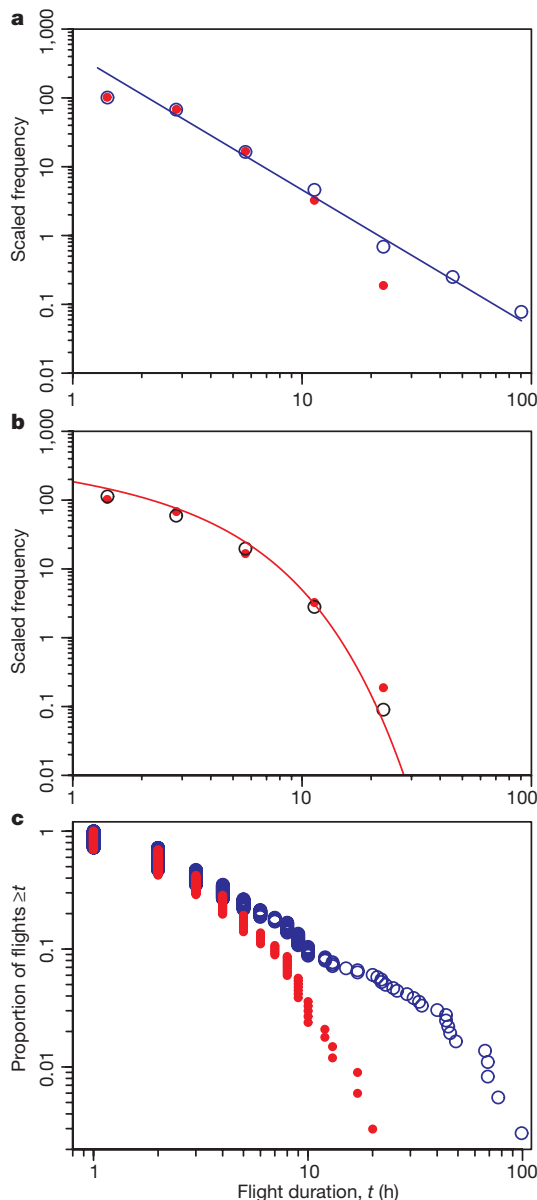
Figure 3 | When corrected, the 1992 wandering albatross flight durations no longer follow a power law. **a**, Blue open circles show the original log–log histogram of 1992 data (Fig. 3a of ref. 7). Breakpoints of bins are at 1, 2, 4, 8, 16, 32, 64 and 128 h (with bin intervals $1 \le t < 2$, $2 \le t < 4$, and so on), and results are plotted at the geometric means. The frequencies are each normalized by their respective bin widths to yield frequency densities that compensate for the increasing bin widths[30] (termed logarithmic binning with normalization, LBN, in ref. 24). The straight line indicates a power law of exponent $\mu = 2$ (ref. 7). Red filled circles are adjusted flight durations that take into account time spent on Bird Island, binned in the same manner, showing no power-law behaviour. **b**, The gamma distribution fitted to the (unbinned) flight durations (red curve) has MLE values $s = 0.73$ (95% CI: 0.19–1.32) and $r = 0.33\,\mathrm{h}^{-1}$ (95% CI: 0.22–0.46), and the data are consistent with coming from this distribution ($n = 335$, degrees of freedom = 8, $G = 11.9$, $P = 0.16$). This distribution yields expected counts in each bin (black open circles), which are what should be compared with the binned data. Our multinomial likelihood approach accounts for the fact that the loggers' memory limitations meant that a record of 1 h could correspond to a flight anywhere in the range 1–3 h. This fact, plus the effects of the binning procedure, result in the differences between the red curve and black circles (note the log scale); see Supplementary Information. **c**, Original (blue open circles) and adjusted (red filled circles) data as a rank/frequency plot. Each record yields a point, and because the resolution of the logger data was 1 h, there can be multiple points for each given flight duration t. The ordinate shows proportion rather than number of flights, because of the different sizes of the data sets.

reported power-law μ values[10] lie outside the 95% confidence intervals (CIs) for the maximum likelihood estimates (MLEs), and are inconsistent with the data. Figure 4 shows the log–log histograms of

---

Box 1 | When is a power law not a power law?

The approach widely used to test for biological Lévy flight search patterns has been: (1) plot the move-length data as some form of histogram on log–log axes; (2) draw or fit a straight line across the full range of data or just the tail; (3) define μ to be the negative of the slope of the line; (4) conclude that the data follow a power law of exponent μ (across the full range or just the tail); (5) then if $1 < \mu \le 3$ conclude that the organism performs a Lévy flight with exponent μ.

It is well known that log–log axes tend to make relationships look straight, and so it is problematic to only plot the data on a log–log plot and then conclude that the data lie on a straight line. One should at least consider an alternative move-length distribution, such as the exponential that corresponds to a simple uncorrelated Poisson random process. Reference 16 did also test the exponential, although used the unreliable[24,25] LT method for the power law and compared the distributions by comparing coefficients of variation[29] (R^2), which is not useful for choosing between models[19].

Here we summarize how to use modern statistical methods of model selection[19,20] to test whether a given data set $\mathbf{x} = \{x_1, x_2, x_3, \ldots, x_n\}$ provides more evidence for a power-law tail or an exponential tail. Considering the tail to start at a, the power-law tail has probability density function

$$f_1(x) = Cx^{-\mu}, \quad x \ge a \tag{3}$$

where the normalization constant $C = (\mu - 1)a^{\mu-1}$, and the exponential tail has probability density function

$$f_2(x) = \lambda e^{-\lambda(x-a)}, \quad x \ge a \tag{4}$$

The log-likelihood function[18,19] for the power law is[23]

$$\log[L_1(\mu|\mathrm{data}\ \mathbf{x})] = n \log(\mu - 1) + n(\mu - 1)\log a - \mu \sum_{j=1}^{n} \log x_j \tag{5}$$

where $L_1(\mu|\mathrm{data}\ \mathbf{x})$ is the likelihood of a particular value of the unknown parameter μ given the known data $\mathbf{x}$ (and log is natural log). For the exponential model the unknown parameter is λ, and

$$\log[L_2(\lambda|\mathrm{data}\ \mathbf{x})] = n \log \lambda + n\lambda a - \lambda \sum_{j=1}^{n} x_j \tag{6}$$

Solving for the MLEs analytically[23] gives $\hat{\mu} = 1 - n / \left(n \log a - \sum_{j=1}^{n} \log x_j \right)$ and $\hat{\lambda} = 1 / \left(\sum_{j=1}^{n} x_j / n - a \right)$. Akaike's information criterion[18,19] (AIC) for model i ($i = 1, 2$) is

$$\mathrm{AIC}_i = -2 \log\left[L_i\left(\hat{\theta}_i | \mathrm{data}\ \mathbf{x} \right) \right] + 2K_i \tag{7}$$

where $\hat{\theta}_1 = \hat{\mu}, \hat{\theta}_2 = \hat{\lambda}$, and K_i is the number of parameters being estimated for model i ($K_1 = K_2$ here). The best model is the one with the minimum AIC, $\mathrm{AIC}_{\min}$. Then, AIC differences are given by $\Delta_i = \mathrm{AIC}_i - \mathrm{AIC}_{\min}$. The Akaike weights[19] are relative likelihoods of each model, given by

$$w_i = \frac{e^{-\Delta_i/2}}{e^{-\Delta_1/2} + e^{-\Delta_2/2}} \tag{8}$$

normalized so the weights sum to 1. The weight w_i is considered as the weight of evidence in favour of model i being the best model for the given data, out of the models considered. Multiple models of varying complexity may also be simultaneously considered[19]. Natural data that follow a power law must be finitely truncated; any approach that neglected this would suffer to some degree (see Supplementary Information). Future work could explore the problem of inference of truncated[3] Lévy flights.

The likelihood approach clarifies what is meant by the 'tail' of the data (permitting goodness-of-fit tests), eliminates binning problems associated with log–log histograms (outlined in ref. 24), and yields 95% CIs (computed using the profile likelihood-ratio test[18]). Furthermore, if the power law is supported by the data, then the MLE for μ is more accurate (A.M.E., manuscript in preparation) than the estimate obtained from the LBN method.

Table 1 | Re-analysis of data sets for which Lévy flight behaviour was concluded[10]

Data set	Reported μ	MLE for μ (95% CI) G, P	MLE for λ (95% CI) G, P	Akaike weights Power law	Akaike weights Exponential
Deer, unfenced	2.0	1.57 (1.29, 1.85) 19.0, 0.0003	0.0264 (0.0214, 0.0318) 4.2, 0.24	0.0006	0.9994
Deer, fenced	2.1	1.44 (1.06, 1.83) 4.87, 0.088	0.0290 (0.0212, 0.0372) 1.08, 0.58	0.05	0.95
Bumblebees, high food	3.5	3.68 (2.14, 5.42) 0.16, 0.69	0.0153 (0.0086, 0.0234) 0.59, 0.44	0.59	0.41
Bumblebees, low food	2.0	2.20 (1.84, 2.60) 11.3, 0.19	0.00609 (0.00497, 0.00732) 13.1, 0.11	0.78	0.22

Tests are over bounded ranges [a, b], with a, b and reported power-law exponents for μ taken from ref. 10. Goodness-of-fit values G and P are for the G-test[29] (n and degrees of freedom for the four scenarios are, respectively, 141 and 3, 112 and 2, 25 and 1, and 129 and 8). For both deer scenarios, the reported μ values[10] lie outside the 95% CIs. For the unfenced deer, the data are inconsistent with coming from a power law at the maximum likelihood estimate (MLE) value (and so are certainly inconsistent with the reported μ, for which we calculate $G = 27.9$, $P < 10^{-5}$). The Akaike weights overwhelmingly favour the exponential. For the fenced deer, the data are consistent with coming from a power law with the MLE exponent of $\mu = 1.44$ (although not with the reported exponent of $\mu = 2.1$, for which $G = 16.7$, $P = 0.0002$). The Akaike weights favour the exponential fairly convincingly. For the high-food bumblebees the data were considered to follow a power law[10], but with the reported μ lying outside the Lévy range $1 < \mu \leq 3$. The CI computed here does overlap the Lévy range, but the Akaike weights favour neither the power law or exponential. For the low-food bumblebees the reported μ lies within the 95% CI and the power law is indeed favoured (using the MLE value for μ), but not convincingly.

the deer data and reported power laws from ref. 10, plus the exponential distributions calculated in Table 1. By eye, the two distributions do not appear as different as concluded statistically. However, the log–log nature of Fig. 4 acts to minimize (to the eye) any differences, and de-emphasizes the fact that there are far more data represented by some points (bins) than others. This illustrates the problems involved with fitting lines (or curves) to histograms plotted on a log–log scale.

For the bumblebee high-food situation, the Akaike weights essentially favour neither model (Table 1). For the low-food situation, the

power law is favoured, but not convincingly. However, we have assumed maximum attainable flight values (b) based simply on the ranges of the power-law lines drawn in ref. 10. This assumption favours preference for the power law, and relaxing it slightly (because it is very unlikely to hold) eliminates the marginal preference for the power law in the low-food bumblebee scenario (see Supplementary Information). Hence, none of the four data sets exhibits evidence for the power law over the exponential.

Furthermore, the deer foraging times[10] actually correspond to times spent cropping and processing food (that is, handling times) at a particular feeding site (pages 608 and 610 in ref. 27), rather than time spent moving between sites. Also, the bumblebee data actually represent the distances between successively visited flower heads, rather than the flight times or distances flown (Fig. 1 in ref. 28). Therefore, neither of these data sets was ideal for testing for Lévy flight search behaviour. Nonetheless, we have re-analysed these data sets to demonstrate the problems with inferring Lévy flight behaviour by fitting (or drawing) a straight line through points on a log–log histogram, not considering alternative distributions, and not performing goodness-of-fit tests. Our approach used here is grounded in the statistical theory of likelihood[18,19], avoids any binning, and does not ignore bins with zero counts[25]. Given these issues relating to the identification of power laws, we query whether the current evidence for Lévy flights would withstand more rigorous statistical analysis. This then raises questions as to when might a foraging animal satisfy the necessary conditions for a Lévy flight to be an optimal[2,10] search strategy.

METHODS SUMMARY

Maximum likelihood estimates were obtained by numerically maximizing the appropriate log-likelihood functions, and 95% CIs were computed using the profile likelihood-ratio test[18]. All goodness-of-fit tests were performed using the G-test (likelihood-ratio test) with Williams's correction[29]. Akaike weights were computed according to ref. 19, with bias-adjusted AIC_c used for high-food bumblebees due to the small sample size. Computations were performed using R, version 2.2.0 (http://www.r-project.org).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Figure 4 | Foraging times of deer previously concluded to demonstrate Lévy flight behaviour[10]. a, Digitized data of foraging times of unfenced deer from Fig. 3 of ref. 27, as plotted in Fig. 3c of ref. 10. Circles are percentages of foraging times in each bin. Bins are each 20-s wide, although the axes are logarithmic (the LT method of ref. 24). The blue line is the power law of exponent $\mu = 2$ from ref. 10, and the red curve is the superior fit of an exponential tail (see Table 1). **b**, As for **a**, but for deer in fenced areas. The blue line is the power law of exponent $\mu = 2.1$ from ref. 10, and the red curve is the superior fit of an exponential tail.

1. Turchin, P. *Quantitative Analysis of Movement: Measuring and Modeling Population Redistribution in Animals and Plants* (Sinauer, Sunderland, Massachusetts, 1998).
2. Shlesinger, M. F. Search research. *Nature* **443**, 281–282 (2006).
3. Shlesinger, M. F., Zaslavsky, G. M. & Frisch, U. (eds) *Lévy Flights and Related Topics in Physics* (Springer, Berlin, 1995).
4. ben-Avraham, D. & Havlin, S. *Diffusion and Reactions in Fractals and Disordered Systems* (Cambridge Univ. Press, Cambridge, 2000).
5. Bardou, F., Bouchaud, J.-P., Aspect, A. & Cohen-Tannoudji, C. *Lévy Statistics and Laser Cooling: How Rare Events Bring Atoms to Rest* (Cambridge Univ. Press, Cambridge, 2002).
6. Ott, A., Bouchaud, J. P., Langevin, D. & Urbach, W. Anomalous diffusion in ''living polymers'': a genuine Lévy flight? *Phys. Rev. Lett.* **65**, 2201–2204 (1990).
7. Viswanathan, G. M. et al. Lévy flight search patterns of wandering albatrosses. *Nature* **381**, 413–415 (1996).
8. Metzler, R. & Klafter, J. The restaurant at the end of the random walk: recent developments in the description of anomalous transport by fractional dynamics. *J. Phys. A: Math. Gen.* **37**, R161–R208 (2004).
9. Klafter, J. & Sokolov, I. M. Anomalous diffusion spreads its wings. *Phys. World* **18**, 29–32 (2005).
10. Viswanathan, G. M. et al. Optimizing the success of random searches. *Nature* **401**, 911–914 (1999).
11. Reynolds, A. M. Scale-free movement patterns arising from olfactory-driven foraging. *Phys. Rev. E* **72**, 041928 (2005).
12. Bénichou, O., Loverdo, C., Moreau, M. & Voituriez, R. Two-dimensional intermittent search processes: An alternative to Lévy flight strategies. *Phys. Rev. E* **74**, 020102(R) (2006).
13. Mårell, A., Ball, J. P. & Hofgaard, A. Foraging and movement paths of female reindeer: insights from fractal analysis, correlated random walks, and Lévy flights. *Can. J. Zool.* **80**, 854–865 (2002).
14. Bartumeus, F., Peters, F., Pueyo, S., Marrasé, C. & Catalan, J. Helical Lévy walks: Adjusting searching statistics to resource availability in microzooplankton. *Proc. Natl Acad. Sci. USA* **100**, 12771–12775 (2003).
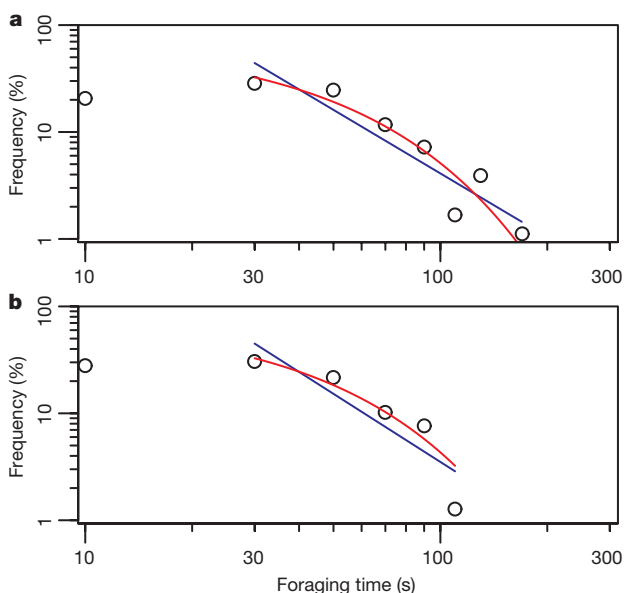
15. Austin, D., Bowen, W. D. & McMillan, J. I. Intraspecific variation in movement patterns: modeling individual behaviour in a large marine predator. *Oikos* **105**, 15–30 (2004).

16. Ramos-Fernández, G. *et al.* Lévy walk patterns in the foraging movements of spider monkeys (*Ateles geoffroyi*). *Behav. Ecol. Sociobiol.* **55**, 223–230 (2004).

17. Bertrand, S., Burgos, J. M., Gerlotto, F. & Atiquipa, J. Lévy trajectories of Peruvian purse-seiners as an indicator of the spatial distribution of anchovy (*Engraulis ringens*). *ICES J. Mar. Sci.* **62**, 477–482 (2005).

18. Hilborn, R. & Mangel, M. *The Ecological Detective: Confronting Models with Data* (Vol. 28, Monographs in Population Biology, Princeton Univ. Press, New Jersey, 1997).

19. Burnham, K. P. & Anderson, D. R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* 2nd edn (Springer, New York, 2002).

20. Johnson, J. B. & Omland, K. S. Model selection in ecology and evolution. *Trends Ecol. Evol.* **19**, 101–108 (2004).

21. Afanasyev, V. A miniature storing activity recorder for seabird species with 80 bytes of memory for data storage. *NERC Tech.* **1**, 4–7 (1993).

22. Shlesinger, M. F. & Klafter, J. Lévy walks versus Lévy flights. In *On Growth and Form: Fractal and Non-Fractal Patterns in Physics* (eds Stanley, H. E. & Ostrowsky, N.) 279–283 (Martinus Nijhoff Publishers, Dordrecht, 1986).

23. Newman, M. E. J. Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.* **46**, 323–351 (2005).

24. Sims, D. W., Righton, D. R. & Pitchford, J. W. Minimizing errors in identifying Lévy flight behaviour of organisms. *J. Anim. Ecol.* **76**, 222–229 (2007).

25. Pueyo, S. & Jovani, R. Comment on ''A keystone mutualism drives pattern in a power function''. *Science* **313**, 1739c (2006).

26. Weimerskirch, H., Gault, A. & Cherel, Y. Prey distribution and patchiness: factors in foraging success and efficiency of wandering albatrosses. *Ecology* **86**, 2611–2622 (2005).

27. Focardi, S., Marcellini, P. & Montanaro, P. Do ungulates exhibit a food density threshold? A field study of optimal foraging and movement patterns. *J. Anim. Ecol.* **65**, 606–620 (1996).

28. Heinrich, B. Resource heterogeneity and patterns of movement in foraging bumblebees. *Oecologia* **40**, 235–245 (1979).

29. Sokal, R. R. & Rohlf, F. J. *Biometry: The Principles and Practice of Statistics in Biological Research* 3rd edn (W. H. Freeman and Company, New York, 1995).

30. Pueyo, S. Diversity: between neutrality and structure. *Oikos* **112**, 392–405 (2006).

Author Contributions A.M.E. performed the analyses, computations and derivations presented in this paper, and led its preparation with input from all authors. The paper presents a synthesis of the work of two different teams of researchers who independently and concurrently conceived of re-examining the Lévy flight hypothesis with newer albatross data. One team (A) comprised N.W.W., M.P.F. and E.J.M. and was subsequently joined by A.M.E. The other team (B) comprised V.A., S.V.B., M.G.E.dL., E.P.R., H.E.S. and G.M.V.; R.A.P. provided albatross expertise to both teams. Team B were the first to show the implication of the long first and last dry sequences on the power-law distribution of albatross flight durations.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.M.E. (EdwardsAnd@pac.dfo-mpo.gc.ca).

METHODS

Initial and final albatross flights. A salt-water logger only detects whether a bird is sitting on the water or not, and its clock starts when it is switched on at a computer. Thus a logger is recording before being attached to a bird, and also, crucially, while the bird sits on a nest[21]. The logger is dry, but the bird is not flying. So the initial sequence of dry readings includes pre-take-off time plus time spent in flight before first landing on water. For the 2004 data we used the GPS data to determine when the birds left Bird Island, and hence obtained the duration of the initial flights (to within an hour), eliminating the time spent on the nest. Return times to the island were used to determine durations of final flights, eliminating the dry period between landing and logger retrieval.

In 1992, PTT devices were used in 6 of the 19 trips (Fig. 2). Departure and return times were calculated to within 2 h (typically 1 h) from a combination of the PTT fixes and direct observations of the birds. For the remaining 13 trips, departure and return times remain unknown. As each trip's initial and final dry sequences were potentially spurious, we omitted them from the data. If the albatrosses really exhibited Lévy flight behaviour then there would be nothing special about the first and last flights, so eliminating these 26 dry sequences should not markedly impact the results. However, the 14 longest dry sequences (for the 13 trips), ranging from 21–99 h, all occurred at the start or end of trips and were eliminated. These were much longer than the maximum flight of 14.9 h in 2004.

Albatross data analysis. The 2004 data consisted of a wet/dry reading every 10 s. A flight is therefore represented by a sequence of consecutive dry readings in between two wet readings. For example, a sequence wet-dry-dry-dry-dry-wet gives a record of 4 dry readings, and represents a flight in the range 30–50 s. We assume 30 s to be the minimum time after take off for a bird to start searching for new food sources. This excludes instances where a bird may have only lifted its leg out of the water to scratch (yielding a sequence wet-dry-wet), or abandoned a take off.

The 1992 data consisted of the number of 15-s intervals within each hour for which a bird was considered wet (for example, Fig. 1 of ref. 7). Flights <1 h could not be directly inferred from the data. In ref. 7, consecutive hourly wet counts of 0 (that is, fully dry hours), in between non-zero hours, were used to infer flight times of 1, 2, 3, … h. A sequence of hours wet-dry-wet was then assumed to be a flight of 1 h. However, such a record could come from a true flight anywhere in the range 1–3 h. In the Supplementary Information we derive the likelihood functions for each data set, taking into account the respective sampling protocols.

Deer and bumblebee data analysis. We digitized the deer and bumblebee data from the original histograms[27, 28] and from ref. 10. For the deer data the original linear histograms[27] were log-transformed[10] (LT method[24]), as reproduced in Fig. 4 from the original data[27], whereas the bumblebee bins[28] were smoothed using running averaging and then lumped to produce log–log histograms[10] (see Supplementary Information). However, our likelihood approach and results in Table 1 do not involve any smoothing or lumping of the original binned data.

We fit the power-law and exponential distributions over unbounded, $[a, \infty)$, and bounded, $[a, b]$, ranges. For the bounded power law we consider the probability density function

$$f(x) = Cx^{-\mu}, \; x \in [a,b] \tag{9}$$

where C is the normalization constant given by $C = (\mu - 1)/(a^{1-\mu} - b^{1-\mu})$. For the bounded exponential distribution, we have

$$f(x) = Ae^{-\lambda x}, \; x \in [a,b] \tag{10}$$

where A is the normalization constant given by $A = \lambda/(e^{-\lambda a} - e^{-\lambda b})$. The data were digitized from histograms[27,28], and so are already binned. In the Supplementary Information we derive the likelihood functions for each distribution, modified from those in Box 1 to account for the pre-binned data and the bounded ranges. For each scenario in Table 1 we took $[a, b]$ to be the ranges implied by the power law straight lines drawn in ref. 10. These ranges imply that it is impossible to obtain data larger than that observed, an assumption that is very unlikely to be true in practice, but that we find favours preference for the power law over the exponential. We investigate sensitivity to b in the Supplementary Information, showing that the result of the power law being slightly favoured for the low-food bumblebees (Table 1) is not robust to relaxation of this assumption.

LETTERS

The amniote primitive streak is defined by epithelial cell intercalation before gastrulation

Octavian Voiculescu[1], Federica Bertocchini[1], Lewis Wolpert[1], Ray E. Keller[2] & Claudio D. Stern[1]

During gastrulation, a single epithelial cell layer, the ectoderm, generates two others: the mesoderm and the endoderm. In amniotes (birds and mammals), mesendoderm formation occurs through an axial midline structure, the primitive streak[1], the formation of which is preceded by massive 'polonaise' movements[2,3] of ectoderm cells. The mechanisms controlling these processes are unknown. Here, using multi-photon time-lapse microscopy of chick (*Gallus gallus*) embryos, we reveal a medio-lateral cell intercalation confined to the ectodermal subdomain where the streak will later form. This intercalation event differs from the convergent extension movements of the mesoderm described in fish and amphibians (anamniotes)[4–8]: it occurs before gastrulation and within a tight columnar epithelium. Fibroblast growth factor from the extra-embryonic endoderm (hypoblast, a cell layer unique to amniotes) directs the expression of Wnt planar-cell-polarity pathway components to the intercalation domain. Disruption of this Wnt pathway causes the mesendoderm to form peripherally, as in anamniotes[1,9]. We propose that the amniote primitive streak evolved from the ancestral blastopore by acquisition of an additional medio-lateral intercalation event, preceding gastrulation and acting independently of mesendoderm formation to position the primitive streak at the midline.

Low magnification time-lapse movies of pre-primitive-streak-stage chick embryos, in which some epiblast cells have been labelled by electroporation of green fluorescent protein (GFP), confirm the 'polonaise movements' described long ago[2,3] (Fig. 1a, b, and Supplementary Movie 1). These are massive, bilaterally symmetrical movements of epiblast cells towards the posterior edge of the disc where the primitive streak will later arise, combined with anterior movement at the midline. Gastrulation then begins by formation of the streak. This arises as a short, broad local thickening of the epiblast, which quickly narrows and elongates towards the centre of the disc.

Three main models have been put forward to account for these movements. One proposes that cell division is increased at the midline and aligned with it to drive streak elongation[10]. The second suggests that long-range attractive and repulsive cues, emanating mainly from the site of primitive streak formation, guide cells to generate the polonaise movements[11,12]. The third proposes that convergent extension is responsible for streak elongation[13]. All three proposals are plausible, but it has been impossible to test them because the behaviour of individual cells could not be visualized directly. Furthermore, the mechanics of cell movements are generally not understood in epithelia such as the chick epiblast, the cells of which (7 μm diameter, 50 μm tall) are connected by apical tight junctions.

We used multi-photon time-lapse microscopy to follow nuclei and cell morphology simultaneously (Fig. 1c). Scattered cells throughout the epiblast were double-labelled by co-electroporation of histone-2B–EGFP (enhanced GFP) and cytoplasmic DsRed-Express and

filmed from stage XII (ref. 14, shortly after the polonaise movements start). A region centred on the site of future primitive streak formation was imaged (Fig. 1d, e) in 15 embryos. The descriptions that follow concentrate on one of these, for clarity. A short broad streak (stage 2, ref. 15) formed after 210 min, which became extended (stage 3) by 320 min.

First, we analysed the frequency and orientation of mitoses. Cell divisions are uniformly distributed across the image field and are not significantly aligned with the midline (Supplementary Movie 2). Rather, cells divide either perpendicularly or along the trajectory of their mother cells, which they tend to continue (either side-by-side or one in front of the other). On average, cells divide every 12.2 h, about twice the time taken for the streak to form. Thus, as has been argued recently[16], neither local accumulation of mitoses nor the orientation of cell division can explain the polonaise movements or the elongation of the primitive streak.

The trajectories of marked nuclei (Fig. 1f–i and Supplementary Movie 3) revealed symmetrical movements with respect to two axes: the future primitive streak axis ('vertical') and an arc parallel to the posterior marginal zone, near the edge of the area pellucida ('horizontal'). Cells converge to the midline and move away from the horizontal axis; the speed of both movements increases with distance from the axes. Close to either axis, cells follow jagged paths, whereas distant cells move more smoothly.

Although not inconsistent with the idea[11,12] that lateral cells are attracted to the midline and then become repelled, these movements can also be explained by medio-lateral cell intercalation (cells wedging themselves between their lateral neighbours), restricted to a small posterior domain. To discriminate between these alternatives, we assessed intercalation by plotting the changes in relative positions of cells in different regions by drawing lines between them ('constellations'; Fig. 1j–m and Supplementary Movie 4). In the central domain, constellations (yellow) change their configuration extensively, becoming narrower and radially aligned (vertically) irrespective of their position and size. In contrast, more lateral cells (purple in Supplementary Movie 4, left) move coherently and seem to be displaced passively. Within the yellow constellations, cell intercalation becomes obvious by observing the cytoplasmic marker (Fig. 1n–q). Thus, as in other cases of intercalation[7,17], all cells in the central domain execute similar movements with respect to their neighbours, but the additive effect of these small local displacements causes more distant cells to move faster. Both in the intercalation region and lateral to it, cells are strongly oriented along the horizontal axis (Fig. 1s) and tend to be bipolar (their leading and trailing protrusions are of comparable size, $P > 0.7$). Importantly, however, cells in the intercalation domain produce significantly larger protrusions than lateral cells (Fig. 1t, $P < 0.0001$). The size of the intercalation domain remains approximately $12 \times 10^4\ \mu m^2$ (3,000 cells) until it starts to

[1]Department of Anatomy & Developmental Biology, University College London, Gower Street, London WC1E 6BT, UK. [2]Department of Biology, Gilmer Hall, University of Virginia, Charlottesville, Virginia 22903, USA.
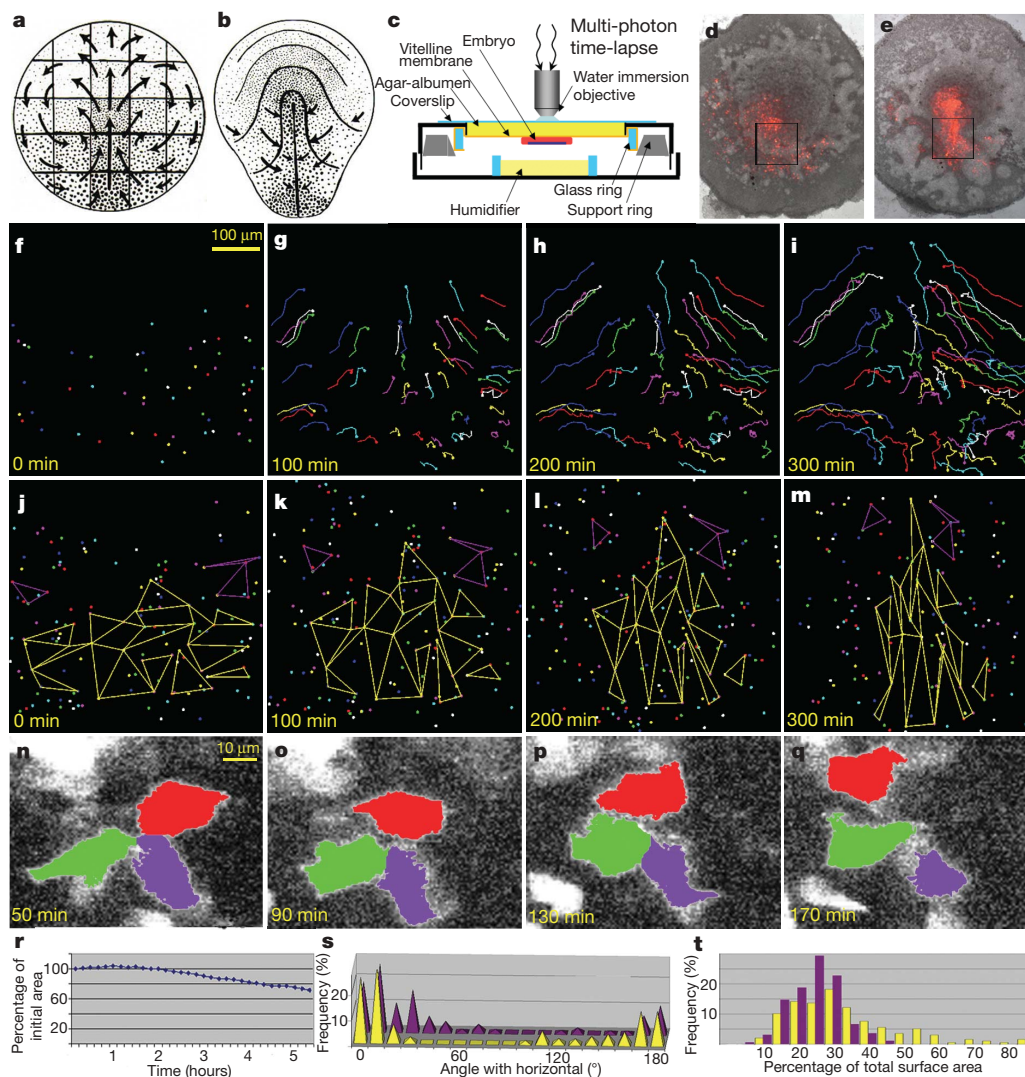
Figure 1 | Cell movements and behaviours during primitive streak formation. a, **b**, Global cell movements in the epiblast, before (**a**) and during (**b**) primitive streak formation (modified from refs. 2, 3). **c**, Culture system for multi-photon time-lapse imaging. **d**, **e**, Imaged embryo before (**d**) and after (**e**) filming (the imaged region is boxed). **f–i**, Subset of the tracked nuclei. **j–m**, Changes in the relative positions of cells over time. In the streak-forming region, cells intercalate extensively (yellow polygons), whereas cells in adjacent regions are displaced passively (purple triangles). **n–q**, Morphology and behaviour of cells in the intercalation domain. **r**, Changes in the area of the intercalation domain over time. **s**, Distribution of the orientation of cell protrusions in the intercalation region (yellow) and lateral to it (purple). **t**, Magnitude of protrusions in the same regions.

decrease around the time of primitive streak formation (stage 2; $t = 3.5$ h in Fig. 1f–m, r and Supplementary Movies 2–4); this indicates that the movements occur in the plane of the epiblast. Initially, all convergence is translated into extension; after the primitive streak appears (stage 2), up to one-third of extension caused by continuing convergence becomes transformed into thickening of the streak.

In the mesoderm of amphibians and teleost fish, cell intercalation depends on the Wnt planar-cell-polarity (Wnt-PCP) pathway[4,18]. A principal ligand activating this pathway is Wnt11 (refs 4, 19). In chick, *WNT11* is uniformly expressed before primitive streak formation[20], but we found three genes in this pathway with localized expression (Fig. 2 and Supplementary Fig. 5): at stages XII–XIV, *FMI1* (flamingo 1) and *PRICKLE1* are concentrated at the posterior edge of the epiblast and in a medial posterior domain; *VANGL2* (vangogh-like2) expression is broader, covering a large part of the inner epiblast, but is especially strong in a small posterior region. Thus, the expression of these genes overlaps in a domain similar to that undergoing intercalation. After stage XIV, the region of overlap changes shape exactly like the intercalation domain (yellow constellations in Fig. 1j–m), suggesting that cells retain their co-expression of these genes as they intercalate.

To test whether the Wnt-PCP pathway is required for streak morphogenesis, we electroporated *Xenopus laevis Dsh* (dishevelled) lacking the DEP domain (*Dsh-ΔDEP*), a specific inhibitor of this pathway[21], into the intercalation region (Fig. 3a). Primitive streak morphogenesis was profoundly compromised (15 out of 18 embryos). In 9 out of 18 embryos, no axial streak developed; instead there was a broad, thickened, *BRA* (brachyury)-expressing crescent at the posterior margin (reminiscent of the shield of teleosts or the blastopore of amphibians) (Fig. 3c, g, h). In another six cases, several
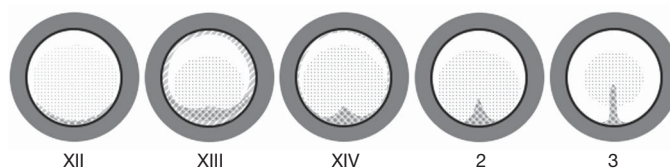

Figure 2 | Summary of expression patterns of Wnt-PCP genes. The diagrams are based on the data shown in Supplementary Fig. 5. Stages are indicated below each diagram. *FMI1*, forward slashes; *PRICKLE1*, reverse slashes and *VANGL2*, stippled.

small, disorganized streaks formed side-by-side in the posterior epiblast (see Supplementary Fig. 6). Targeting as few as one-third to one-half of cells within the future primitive streak domain is
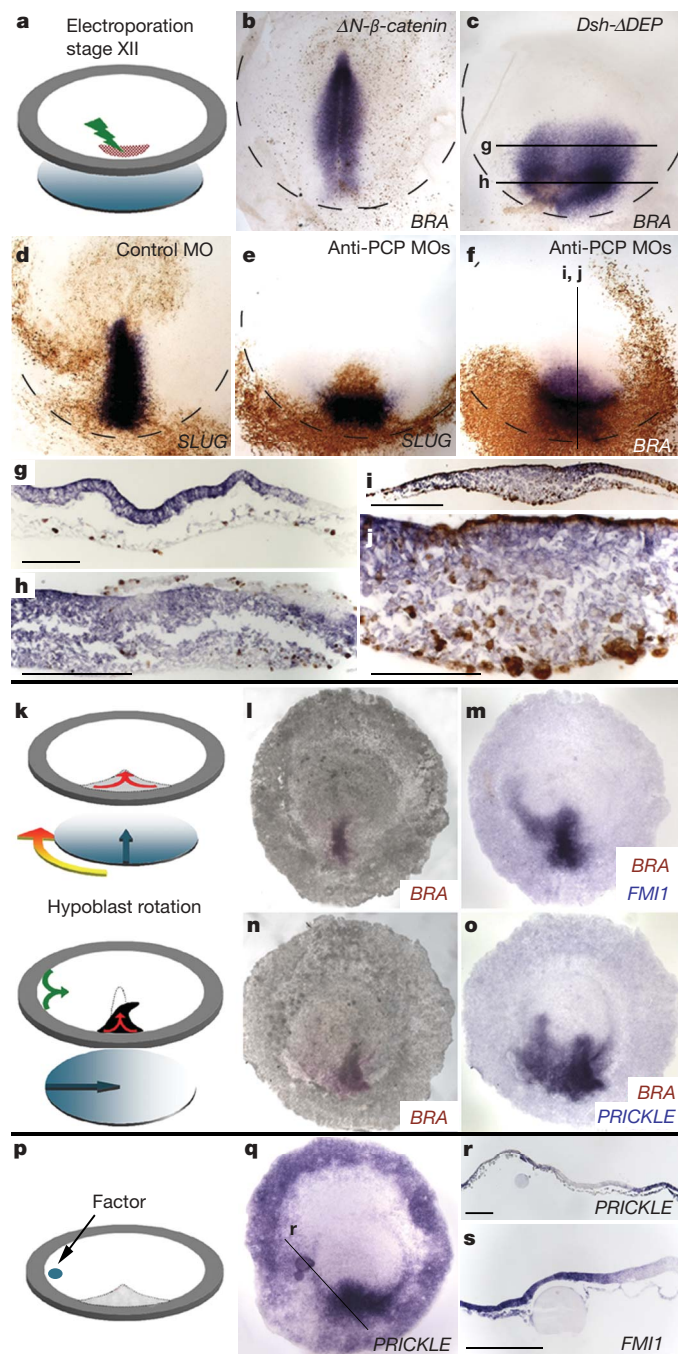

Figure 3 | The Wnt-PCP pathway controls primitive streak morphogenesis.
a–j, Inhibiting the Wnt-PCP pathway blocks cell intercalation, but not mesendoderm formation. **a,** Experimental setup. Electroporation of stabilized β-catenin has no effect (**b**), whereas inhibition of the Wnt-PCP pathway by *Dsh-ΔDEP* prevents formation of a radial streak; gastrulation now occurs in a peripheral crescent (**c**). Morpholino (MO)-mediated knock-down of chick *FMI1*, *VANGL2* and *PRICKLE1* has the same effect, revealed by *SLUG* (**d, e**) and *BRA* (**f**). Transverse sections from *Dsh-ΔDEP* embryos show normal ingression of mesendoderm (**g, h**). Longitudinal sections from MO-embryos (**i, j**) show normal ingression of electroporated cells. **k–o,** Hypoblast rotation induces ectopic expression of Wnt-PCP genes. **k,** Experimental design. *FMI1* (**m,** purple) is induced, but not *BRA* (**l;** same embryo as in **m**). *PRICKLE1* (**o**) is induced, but not *BRA* (**n;** same embryo as **o**). **p–s,** Fgf8 induces both *FMI1* and *PRICKLE1*. **p,** Experimental design. **q, r,** Within 6 h, Fgf8-soaked beads induce *PRICKLE1* (purple) (**q, r**) and *FMI1* (**s**). Scale bars, 100 μm (**g, h, j, s**); 200 μm (**i, r**).

sufficient to cause these defects, even though the construct should act cell-autonomously. This suggests that effective medio-lateral intercalation requires activation of the Wnt-PCP pathway in all neighbouring intercalating cells. The effect cannot be caused by activation of the canonical Wnt pathway, because embryos electro-porated with stabilized β-catenin[22] were indistinguishable from EGFP-electroporated controls (Fig. 3b). To confirm this conclusion and to test the role of the three Wnt-PCP genes that show localized expression, we co-electroporated morpholinos against *FMI1*, *VANGL2* and *PRICKLE1*. In 16 of 35 cases, this also resulted in a marginal crescent of thickened epiblast, expressing the streak mar-kers *BRA* and *SLUG* (snail homologue 2, also known as *SNAI2*) (Fig. 3e, f, i, j). Electroporation of these morpholinos individually or in pairs did not produce defects, nor did a control morpholino (Fig. 3d; *n* = 18 for each). An important observation from both loss-of-function experiments is that ingression of the targeted cells is not affected (Fig. 3g–j); about half of them ingress (as in controls).

What mechanisms are responsible for positioning the site of cell intercalation? The hypoblast (an amniote-specific extra-embryonic tissue, equivalent to the mouse anterior visceral endoderm[23,24]) underlying the ectoderm has long been known to influence axis formation. It does this in two separate ways. It restricts the site and timing of mesendoderm induction by modulating Nodal activity[24]. It also influences epiblast cell movements: 90° rotation of the hypoblast induces a new set of polonaise movements, leading to bending of the streak[23,25] (Fig. 3k). Could the hypoblast regulate cell movements by controlling the expression of Wnt-PCP genes? Rotation of the hypo-blast by 90° quickly (<6 h) induces a new expression domain of *FMI1* and *PRICKLE1* in the epiblast adjacent to the rotated hypoblast
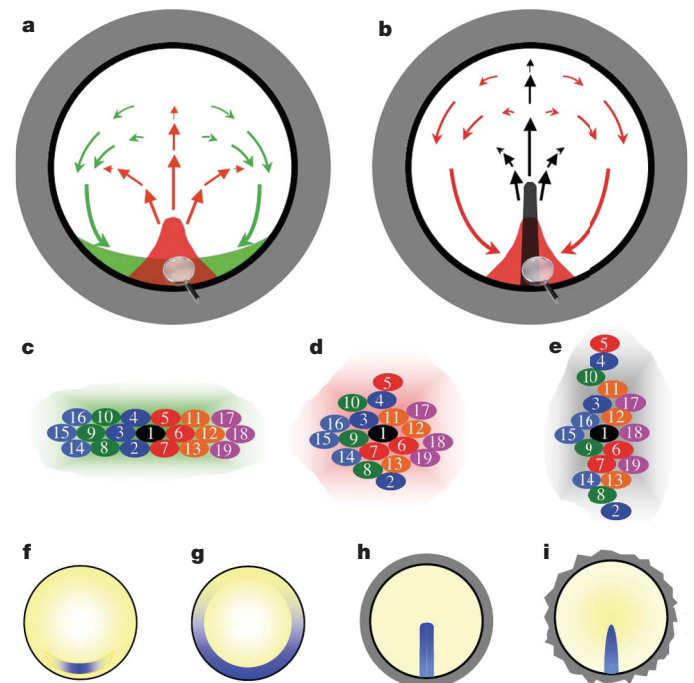

Figure 4 | Early medio-lateral intercalation as an evolutionary step that shapes the primitive streak. a–e, Cell rearrangements driving streak formation. Medio-lateral intercalation changes the shape of the streak-forming region, shown at stage XII (green in **a**), stage 2 (red in **a** and **b**) and stage 3 (grey in **b**). Intercalation displaces lateral epiblast cells towards the streak (green arrows in **a**; red arrows in **b**) and forwards in front of it (red arrows in **a**; black arrows in **b**). **c–e,** The regions shown in **a** and **b** at higher magnification, showing the intercalating cells. **f–i,** Evolution of the ingression site (blue), showing embryos of *Xenopus* (stage 10) viewed from the animal pole (**f**), zebrafish (50% epiboly) viewed from the animal pole (**g**), chick in dorsal view (**h**) and mouse (embryonic day 6.75) viewed from the proximal end (**i**).

(Fig. 3m, o). In contrast, *BRA* is weakly or not induced (Fig. 3l, n), indicating that the hypoblast regulates expression of these Wnt-PCP pathway components independently of mesoderm formation. To identify the signals involved in this induction, we tested the inducing ability of several factors known to be secreted by the hypoblast (dickkopf homologue 1 (Dkk1), crescent, cerberus1 (Cer1), fibroblast growth factor 8 (Fgf8) and combinations thereof) (Fig. 3p). Only Fgf8 upregulates both *FMI1* (17 out of 21) and *PRICKLE1* (17 out of 19) but not *BRA* in the epiblast within 6 h (Fig. 3q–s). These findings suggest that localization of Wnt-PCP activity is controlled, at least partly, at the transcriptional level.

We propose that local intercalation in the epiblast is responsible for positioning and shaping the primitive streak and can also explain the polonaise movements (see Fig. 4a–e) without the need for long-range gradients[11,12]. Convergent extension of the axial mesoderm and neural plate in anamniotes[5–8] is almost certainly conserved in amniotes[26–28], but our study reveals an additional, much earlier (pregastrula) cell intercalation, required for morphogenesis of the primitive streak independently of mesendoderm specification. This is apparently unique to amniotes (Fig. 4f–i) and provides a possible answer to the classical question[9,29,30] of how evolution converted the equatorial blastopore or shield of Anamnia into the radially oriented primitive streak of amniotes.

METHODS SUMMARY

Electroporation was carried out on explanted embryos using five 50 ms pulses of 4.2 V, 500 ms apart, after delivering DNA (total 2 mg ml^{-1}) or morpholinos (total 2.4 mM, equimolar concentrations). Imaging chambers were constructed as shown in Fig. 1, and embryos were filmed with a Leica SP2 microscope with a Tsunami infrared laser tuned to 910 nm. *In situ* hybridization, immunostaining and factor delivery followed standard methods. For full methods, see Supplementary Information.

1. Stern, C. D. (ed.) *Gastrulation: from cells to embryo* (Cold Spring Harbor Laboratory Press, New York, 2004).
2. Gräper, L. Die Primitiventwicklung des Hünchens nach stereokinematographischen Untersuchungen, kontrolliert durch vitale Farbmarkierung und verglichen mit der Entwicklung anderer Wierbeltiere. *Arch. EntwMech. Org.* **116**, 382–429 (1929).
3. Wetzel, R. Untersuchungen am Hühnchen. Die Entwicklung des Keims während der ersten beiden Bruttage. *Arch. EntwMech. Org.* **119**, 188–321 (1929).
4. Heisenberg, C. P. *et al.* Silberblick/Wnt11 mediates convergent extension movements during zebrafish gastrulation. *Nature* **405**, 76–81 (2000).
5. Tada, M., Concha, M. L. & Heisenberg, C. P. Non-canonical Wnt signalling and regulation of gastrulation movements. *Semin. Cell Dev. Biol.* **13**, 251–260 (2002).
6. Wallingford, J. B., Fraser, S. E. & Harland, R. M. Convergent extension: the molecular control of polarized cell movement during embryonic development. *Dev. Cell* **2**, 695–706 (2002).
7. Keller, R. & Davidson, L. in *Gastrulation: from cells to embryo* (ed. Stern, C. D.) 291–304 (Cold Spring Harbor Laboratory Press, New York, 2004).
8. Solnica-Krezel, L. Conserved patterns of cell movements during vertebrate gastrulation. *Curr. Biol.* **15**, R213–R228 (2005).
9. Arendt, D. & Nubler-Jung, K. Rearranging gastrulation in the name of yolk: evolution of gastrulation in yolk-rich amniote eggs. *Mech. Dev.* **81**, 3–22 (1999).
10. Wei, Y. & Mikawa, T. Formation of the avian primitive streak from spatially restricted blastoderm: evidence for polarized cell division in the elongating streak. *Development* **127**, 87–96 (2000).
11. Cui, C., Yang, X., Chuai, M., Glazier, J. A. & Weijer, C. J. Analysis of tissue flow patterns during primitive streak formation in the chick embryo. *Dev. Biol.* **284**, 37–47 (2005).
12. Chuai, M. *et al.* Cell movement during chick primitive streak formation. *Dev. Biol.* **296**, 137–149 (2006).
13. Lawson, A. & Schoenwolf, G. C. Cell populations and morphogenetic movements underlying formation of the avian primitive streak and organizer. *Genesis* **29**, 188–195 (2001).
14. Eyal-Giladi, H. & Kochav, S. From cleavage to primitive streak formation: a complementary normal table and a new look at the first stages of the development of the chick. I. General morphology. *Dev. Biol.* **49**, 321–337 (1976).
15. Hamburger, V. & Hamilton, H. L. A series of normal stages in the development of the chick embryo. *J. Morphol.* **88**, 49–92 (1951).
16. Bodenstein, L. & Stern, C. D. Formation of the chick primitive streak as studied in computer simulations. *J. Theor. Biol.* **233**, 253–269 (2005).
17. Glickman, N. S., Kimmel, C. B., Jones, M. A. & Adams, R. J. Shaping the zebrafish notochord. *Development* **130**, 873–887 (2003).
18. Klein, T. J. & Mlodzik, M. Planar cell polarization: an emerging model points in the right direction. *Annu. Rev. Cell Dev. Biol.* **21**, 155–176 (2005).
19. Medina, A., Reintsch, W. & Steinbeisser, H. *Xenopus* frizzled 7 can act in canonical and non-canonical Wnt signaling pathways: implications on early patterning and morphogenesis. *Mech. Dev.* **92**, 227–237 (2000).
20. Skromne, I. & Stern, C. D. Interactions between Wnt and Vg1 signalling pathways initiate primitive streak formation in the chick embryo. *Development* **128**, 2915–2927 (2001).
21. Rothbächer, U. *et al.* Dishevelled phosphorylation, subcellular localization and multimerization regulate its role in early embryogenesis. *EMBO J.* **19**, 1010–1022 (2000).
22. Domingos, P. M. *et al.* The Wnt/β-catenin pathway posteriorizes neural tissue in *Xenopus* by an indirect mechanism requiring FGF signalling. *Dev. Biol.* **239**, 148–160 (2001).
23. Foley, A. C., Skromne, I. & Stern, C. D. Reconciling different models of forebrain induction and patterning: a dual role for the hypoblast. *Development* **127**, 3839–3854 (2000).
24. Bertocchini, F. & Stern, C. D. The hypoblast of the chick embryo positions the primitive streak by antagonizing nodal signaling. *Dev. Cell* **3**, 735–744 (2002).
25. Waddington, C. H. Experiments on the development of chick and duck embryos cultivated *in vitro*. *Phil. Trans. R. Soc. Lond. B* **221**, 179–230 (1932).
26. Schoenwolf, G. C. & Smith, J. L. Mechanisms of neurulation: traditional viewpoint and recent advances. *Development* **109**, 243–270 (1990).
27. Wang, J. *et al.* Dishevelled genes mediate a conserved mammalian PCP pathway to regulate convergent extension during neurulation. *Development* **133**, 1767–1778 (2006).
28. Ybot-Gonzalez, P. *et al.* Convergent extension, planar-cell-polarity signalling and initiation of mouse neural tube closure. *Development* **134**, 789–799 (2007).
29. Hertwig, O. *Lehrbuch der Entwicklungsgeschichte des Menschen und der Wirbeltiere* (Verlag Gustav Fischer, Jena, 1910).
30. Pasteels, J. Un aperçu comparitif de la gastrulation chez les Chordés. *Biol. Rev. Camb. Philos. Soc.* **15**, 59–106 (1940).

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to O.V. (o.voiculescu@ucl.ac.uk) or C.D.S. (c.stern@ucl.ac.uk).

LETTERS

PLETHORA proteins as dose-dependent master regulators of *Arabidopsis* root development

Carla Galinha[1]*†, Hugo Hofhuis[1]*, Marijn Luijten[1], Viola Willemsen[1], Ikram Blilou[1], Renze Heidstra[1] & Ben Scheres[1]

Factors with a graded distribution can program fields of cells in a dose-dependent manner[1,2], but no evidence has hitherto surfaced for such mechanisms in plants. In the *Arabidopsis thaliana* root, two *PLETHORA* (*PLT*) genes encoding AP2-domain transcription factors have been shown to maintain the activity of stem cells[3]. Here we show that a clade of four *PLT* homologues is necessary for root formation. Promoter activity and protein fusions of PLT homologues display gradient distributions with maxima in the stem cell area. PLT activities are largely additive and dosage dependent. High levels of PLT activity promote stem cell identity and maintenance; lower levels promote mitotic activity of stem cell daughters; and further reduction in levels is required for cell differentiation. Our findings indicate that PLT protein dosage is translated into distinct cellular responses.

During animal development, instructive molecules acquire a graded distribution and induce distinct cellular responses in a concentration-dependent manner. Whether similar mechanisms occur in plants has been controversial; dosage-sensitive action of plant hormones has been inferred only after external application[4]. Plant stem cell regions, which supply cells for the growing root and shoot systems[5], are potential sites of action for instructive gradients. Stem cells are maintained in local micro-environments, which are similar to animal stem cell niches[6]. Stem cell daughters undergo additional divisions in transit-amplifying cell compartments called meristems; when cells leave the meristem they rapidly expand and differentiate. The *PLETHORA1* (*PLT1*, At3g20840) and *PLT2* (At1g51190) genes encode AP2-domain transcription factor family members essential for defining the root stem cell niche[3]. *plt1;plt2* mutants display stem cell loss, loss of transit-amplifying cells and reduced cell expansion. PLT1 and PLT2 expression strongly correlates with a transcriptional response maximum to the plant hormone auxin in the root tip[3,7] and this maximum has been shown to have profound organizing activity[8]—a property often associated with sources of instructive gradients. Here, we reveal that the *PLT* gene family controls distinct aspects of root development in a dose-dependent manner through PLT expression gradients that culminate in the stem cell niche.

The proteins encoded by At5g10510/*AINTEGUMENTA-LIKE6* (*AIL6*)/*PLT3* and At5g17430/*BABY BOOM*(*BBM*) group with PLT1 and PLT2 in the AP2/ERF transcription factor family (Supplementary Fig. 1)[9], and these candidate redundant factors are predicted to be expressed in the root[10].

From the heart-stage of embryogenesis onward, *PLT3* is expressed in provascular cells, the quiescent centre and columella progenitor cells (Fig. 1a). Post-embryonically, *PLT3* messenger RNA accumulates in the root stem cell niche with the strongest signal in the columella stem cell layer (Fig. 1b), in contrast to the predominant quiescent-centre-localization of *PLT1* and *PLT2* transcript[3]. At the heart-stage of embryo development, *BBM* is expressed in provascular

cells and in the lens-shaped quiescent centre progenitor cell (Fig. 1c). Post-embryonically, *BBM* transcript accumulates in the quiescent centre and columella stem cells—in a similar manner to the *PLT* mRNAs—and in provascular cells of the proximal meristem (Fig. 1d).

The *plt3-1* mutant allele carries a T-DNA insertion interrupting the first AP2 domain (Supplementary Fig. 2). No transcript was detected by PCR with reverse transcription (RT–PCR) or by *in situ* hybridization on *plt3-1* seedlings (data not shown), suggesting that *plt3-1* is a null allele. Homozygous *plt3* single mutants have slightly shorter roots and meristems compared to wild type, but $plt1^{-/-}$ $plt2^{-/-}$ $plt3^{-/-}$ triple homozygotes are rootless (Fig. 1e, upper inset). Progeny from $plt1^{-/-}$ $plt2^{-/-}$ $plt3^{+/-}$, $plt1^{-/-}$ $plt2^{+/-}$ $plt3^{-/-}$ and $plt1^{+/-}$ $plt2^{-/-}$ $plt3^{-/-}$ plants segregate ~25% rootless triple mutants (Supplementary Table 2), demonstrating linkage between the rootless phenotype and the three *PLT* genes. The embryonic root pole of triple homozygous seedlings is fully differentiated at 3 days post germination (d.p.g.) and adventitious root primordia arrest at 6 d.p.g. (Supplementary Fig. 3). Mature $plt1^{-/-}$ $plt2^{-/-}$ embryos have only subtle defects in the cellular organization of the distal-most region[3] (Supplementary Fig. 4), but $plt1^{+/-}$ $plt2^{-/-}$ $plt3^{-/-}$ parents yield ~25% embryos with aberrant root poles that lack a lateral root cap cell layer (Supplementary Fig. 4).

We previously showed that $plt1^{-/-}$ $plt2^{-/-}$ mutants have strongly reduced transcription of the *PIN4* gene, which encodes an auxin efflux facilitator[11]. In triple mutant embryos from $plt1^{+/-}$ $plt2^{-/-}$ $plt3^{-/-}$ parents, *PIN1* and *PIN3* mRNAs are strongly reduced (Fig. 1g–j and Supplementary Table 1). Post-embryonic *PIN2* mRNA is strongly reduced in triple mutant roots before differentiation (Fig. 1k, l). Therefore, PLT1, PLT2 and PLT3 redundantly control expression of multiple *PIN* genes in the embryonic and post-embryonic root.

bbm-1 and *bbm-2* mutant alleles carry T-DNA insertions before and in the beginning of the first AP2 domain, respectively (Supplementary Fig. 2). Truncated transcripts are detected by RT–PCR and may be translated, but genetic interactions (described below) suggest that the insertions cause loss-of-function effects. $plt3^{-/-}$ $bbm^{-/-}$ double mutants have a shorter root and root meristem than either single mutant (Fig. 1f, and Supplementary Fig. 3).

Intriguingly, the progeny of plants segregating different *plt* and *bbm* allele combinations lack root and hypocotyl (Fig. 1e, lower inset) at significant frequencies (Supplementary Table 2), reaching ~10% of the progeny of selfed $plt1^{-/-}$ $plt2^{+/-}$ $plt3^{-/-}$ $bbm-2^{-/-}$. These defects initiate in the early basal embryo (Supplementary Fig. 5) and resemble those in mutants of the auxin response factor *MONOPTEROS*[12] and the auxin perception machinery[13,14]. *PLT* genes do not seem to strongly perturb early global auxin-dependent patterning processes, as suggested by essentially normal cotyledon vasculature in the triple mutant (Supplementary Fig. 4). Segregation

[1]Department of Biology, Faculty of Science, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands. †Present address: Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RB, UK.
*These authors contributed equally to this work.

of *plt2* in a homozygous *bbm* background and vice versa yields ~25% early arrested embryos, and homozygous double mutants could not be recovered, indicating a redundant function in early embryogenesis (data not shown).

Ectopic root structures are initiated by constitutive embryonic expression of *PLT* genes[3] and after induction of *BBM* expression[15]. To test whether PLT induction induces a developmental switch to root development, we expressed a PLT2–GR fusion protein that complements $plt1^{-/-} plt2^{-/-}$ after dexamethasone (dex) induction, when driven by its own promoter (Supplementary Fig. 6). When *35S-PLT2–GR* is activated by application of dex, roots are produced from the shoot apex (Fig. 1m, n). Our gain- and loss-of-function experiments indicate that *PLT* genes are master switches for root development.

$plt1^{-/-} plt2^{+/-} plt3^{-/-}$ mutants have intermediate root and meristem size between $plt1^{-/-} plt3^{-/-}$ and $plt1^{-/-} plt2^{-/-} plt3^{-/-}$ (Fig. 1e, f, and Supplementary Fig. 3) and 50% of $plt1^{-/-} plt2^{+/-} plt3^{-/-} bbm^{-/-}$ seedlings have shorter roots than $plt1^{-/-} plt3^{-/-} bbm^{-/-}$, whereas 50% have no primary root (Supplementary Table 2). *plt3* alleles are also semi-dominant, because growth and meristem maintenance defects in $plt1^{-/-} plt2^{-/-} plt3^{+/-}$ seedlings are intermediate between $plt1^{-/-} plt2^{-/-}$ and $plt1^{-/-} plt2^{-/-} plt3^{-/-}$ (Fig. 1f, and Supplementary Fig. 3). The semi-dominance of *plt2* and *plt3* loss-of-function alleles indicates dose-dependent activity.

To test whether *PLT* genes equally contribute to PLT 'dosage', we transformed $plt1^{-/-} plt2^{-/-}$ double mutants with *PLT1*, *PLT2*, *PLT3* and *BBM* genes fused to the yellow fluorescent protein gene *YFP* and driven by the full *PLT2* promoter. In independent lines with similar overall YFP levels, *PLT1* and *PLT2* fully complemented and *PLT3* and *BBM* partially complemented root growth in the double mutant. All PLT proteins rescued columella stem cell activity (Supplementary Figs 6 and 7). Thus, total PLT levels and to some extent intrinsic differences in PLT protein activity contribute to root growth and stem cell maintenance.

Transgenic lines carrying complete promoters of the *PLT* genes fused to the cyan fluorescent protein gene *CFP* reveal highest promoter activity in the stem cell niche, consistent with mRNA levels, but they also show graded activity in the proximal meristem (Fig. 2a–d). Gradients can be observed in epidermal surface views, excluding quenching effects, and they are specific to *PLT* promoters (Fig. 2e). To analyse whether this promoter activity drives a PLT protein gradient, we combined the PLT–YFP fusions with their corresponding full promoters. *PLT1* and *PLT2* gene fusions complemented $plt1^{-/-} plt2^{-/-}$ mutants (Supplementary Figs 6 and 7, and data not shown).

Figure 1 | Four *PLT* genes promote root formation. a–d, *In situ* hybridization with *PLT3*- (**a**, **b**) and *BBM*- (**c**, **d**) specific probes in wild-type embryos at heart-stage (**a**, **c**), and in roots of 3 d.p.g. wild-type plants (**b**, **d**). Asterisk, quiescent centre; pv, provasculature; col, columella. **e**, Seedlings 10 d.p.g., from left to right: wild type, $plt3^{-/-}$, $bbm-1^{-/-}$, $plt1^{-/-} plt2^{-/-}$, $plt1^{-/-} plt2^{-/-} plt3^{-/-}$ and a $plt1^{-/-} plt2^{+/-} plt3^{-/-} bbm-1^{-/-}$ segregant. Insets show magnification of $plt1^{-/-} plt2^{-/-} plt3^{-/-}$ mutant (upper) and $plt1^{-/-} plt2^{+/-} bbm-1^{-/-}$ segregant (lower). **f**, Meristem size in wild type (Col0 and WS) and *plt* mutants at the indicated d.p.g. For each data point, $n = 10$ to 50; error bars, s.e.m. **g–l**, *In situ* hybridization using PIN probes on wild-type (**g**, **i**) and $plt1^{-/-} plt2^{-/-} plt3^{-/-}$ (**h**, **j**) torpedo-stage embryos and wild-type (**k**) and $plt1^{-/-} plt2^{-/-} plt3^{-/-}$ mutant (**l**) 2 d.p.g. seedlings. **m**, Shoot of 9 d.p.g. *35S-PLT2–GR* plant 6 days after dexamethasone application. **n**, Magnification reveals cellular organization of ectopic root including columella starch granules.
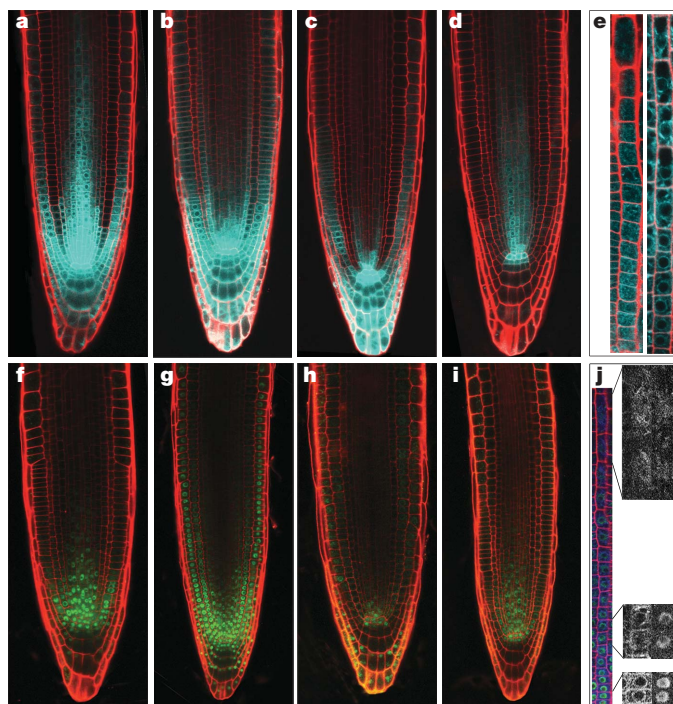

Figure 2 | *PLT* promoter activity and PLT protein fusions display gradients. a–d, CFP reporter driven by full-size promoters of *PLT1* (**a**) *PLT2* (**b**) *PLT3* (**c**) and *BBM* (**d**). **e**, Epidermal gradient of *PLT2* (left) but not *RCH2* (right) promoter. **f–i**, YFP reporter fused in-frame to genomic fragments of *PLT1* (**f**), *PLT2* (**g**), *PLT3* (**h**) and *BBM* (**i**). **j**, Co-localization in one plant of *PLT2* transcriptional (CFP, left magnification) and translational (YFP, right magnification) fusion viewed in different regions using separate channels.

All PLT protein fusions revealed conspicuous gradients that extend into the transit-amplifying cells and, for the PLT2 and PLT3 fusions, into the elongation zone (Fig. 2f–i). The promoter and protein gradients fully match when combined in one plant (Fig. 2j). We previously reported accumulation of *PLT* transcripts in the stem cell area[3], but, after extended staining, *PLT1 in situ* hybridizations also reveal a broader expression domain (Supplementary Fig. 8). We concluded that *PLT* promoter activity leads to protein gradients with maximum expression in the stem cell niche. PLT1 and PLT2 expression maxima broadly encompass the niche, whereas PLT3 and BBM are more restricted.

We asked whether differences in PLT expression domains affect the ability of PLT proteins to compensate for redundant partners. Indeed, *PLT1* and *PLT2* only partially complement a $plt1^{-/-} plt2^{-/-}$ mutant when driven by the *BBM* promoter (Supplementary Figs 6 and 7).

Our experiments suggested that the PLT protein concentration gradient instructs different outputs in different regions, even though each gene slightly differs in activity and expression profile. We therefore tested whether altering the level or shape of the PLT2 gradient affects the position of developmental boundaries. We expressed the PLT2–YFP fusion in $plt1^{-/-} plt2^{-/-}$ mutants under the *RCH2* promoter, which has low activity in the stem cell area but is active in meristematic and elongating cells at a level comparable to that of the *PLT2* promoter (Fig. 3d–g). *RCH2-PLT2–YFP* prolongs transit-amplifying cell divisions but fails to maintain stem cells at 7 d.p.g. (Fig. 3b, d). The transit-amplifying cell pool is lost at 12 d.p.g. (Fig. 3c). We concluded that intermediate PLT levels in the meristem promote transient cell cycling.

To validate that meristem size is controlled by a PLT gradient, we analysed $plt1^{-/-} plt2^{-/-}$ mutants complemented with the PLT2–YFP construct driven by a truncated 1.3-kb *PLT2* promoter fragment (*pPLT2s*). This truncated promoter drives significant expression in the stem cell area but the gradient declines more rapidly (Fig. 3e, f). Accordingly, stem cells are rescued but root and meristem sizes are ~50% smaller (Supplementary Fig. 7). The amount of YFP signal per mid-nuclear section in the stem cell zone, halfway the meristem, and in the first expanding cells, provides three clearly separated intensity ranges that match with zonation in the full- and truncated-promoter driven gradients (Fig. 3g), suggesting that the PLT2 gradient defines meristem zonation.

A dose-dependent gradient model predicts that PLT overexpression shifts the meristem boundary. Indeed, dex induction of *35S-PLT2–GR* plants promotes continuous growth of the transit-amplifying cell pool and meristem size increases (Fig. 4a–c). Ink toner marks marking the elongation zone boundaries at the time of induction reveal that PLT overexpression sustains cell division only in cells that are still cycling and inhibits cell expansion in the elongation zone. These data reinforce the idea that distinct PLT levels dictate cell proliferation and mitotic exit.

The auxin response marker *DR5-GUS*[8] and *PIN3* transcription do not change in *35S-PLT2–GR* plants just before the onset of meristem size expansion, but only at later stages, indicating that PLT-induced expansion of the division zone is not caused by rapid changes in *PIN* expression (Supplementary Fig. 9).

Notably, the stem cell area in PLT2–GR plants is not altered after induction (Fig. 4c). The *RETINOBLASTOMA* (*RBR*) pathway was recently identified as an independent stem cell input[16], so we reasoned that this pathway might still limit stem cell pool size in the presence of higher PLT levels. Therefore, we combined a root-specific RNA interference (RNAi)-mediated silencing construct (*RCH1-RBRi*)[16] with *35S-PLT2–GR*. After induction with dex in the double transgenic, root meristem size increases as in *35S-PLT2–GR*, but clusters of dividing cells in the root cap area expand beyond that seen in *RCH1-RBRi* alone (Fig. 4d, e). Moreover, periclinal divisions normally associated with stem cells occur throughout the proximal area (Fig. 4f). These data suggested that the high expression region of

the PLT gradient can be instructive for stem cell fate. Dramatic support for this notion is provided by duplications of the distal stem cell area in ~10% of *RCH1-RBRi;35S-PLT2–GR* root meristem zones (Fig. 4g–i). We concluded that high PLT levels define the stem cell domain, confirming PLT dosage-dependent stem cell specification.
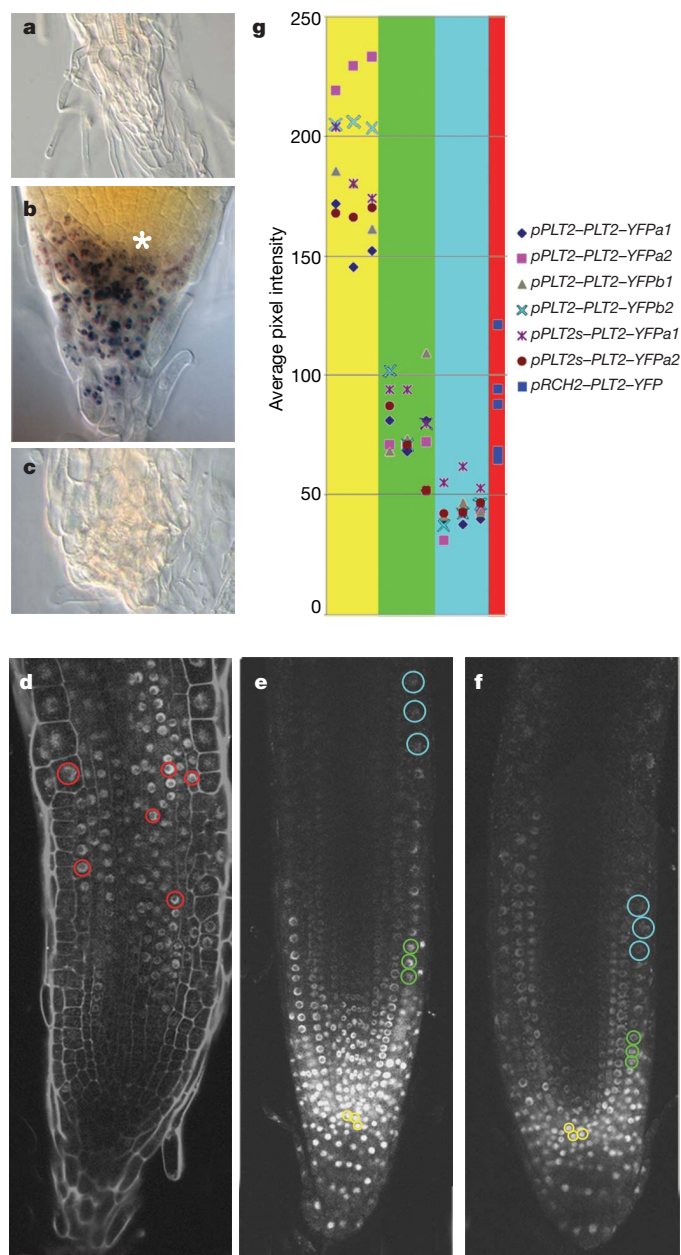

Figure 3 | **PLT expression regulates stem cell maintenance and meristem boundary. a–d,** Meristem prolongation but not stem cell rescue in *RCH2-PLT2–YFP* plants. Nomarski optics image of root tip of 7 d.p.g. *plt1;plt2* (**a**), and of *plt1;plt2 RCH2-PLT2-YFP* at 7 d.p.g. (**b**) and 12 d.p.g. (**c**). Starch granule staining (brown) shows no rescue of columella stem cells below the quiescent centre. Confocal view of 7 d.p.g. *plt1;plt2 RCH2-PLT2–YFP* root (**d**) shows that the meristem is rescued and reveals no expression of PLT2–YFP in the stem cell area. Asterisk in **b**, quiescent centre. **e, f,** Promoter truncation shifts the meristem boundary. CLSM views at identical pinhole and laser settings for *RCH2-PLT2–YFP* (**d**), *pPLT2-PLT2–YFP* (**e**) and *pPLT2s-PLT2–YFP* (**f**). **g,** Quantification of fluorescence per nucleus in *pRCH2-PLT2–YFP* transient meristem (red circles in **d**, and red graph areas), and in stem cells (yellow in **e, f** and graph area), mid-meristem (green in **e, f** and graph area) and first elongating cells (blue in **e, f** and graph areas) of *pPLT2-PLT2–YFP* and pPLT2s-PLT2–YFP (a and b indicate independent transformants, 1 and 2 indicate different roots).
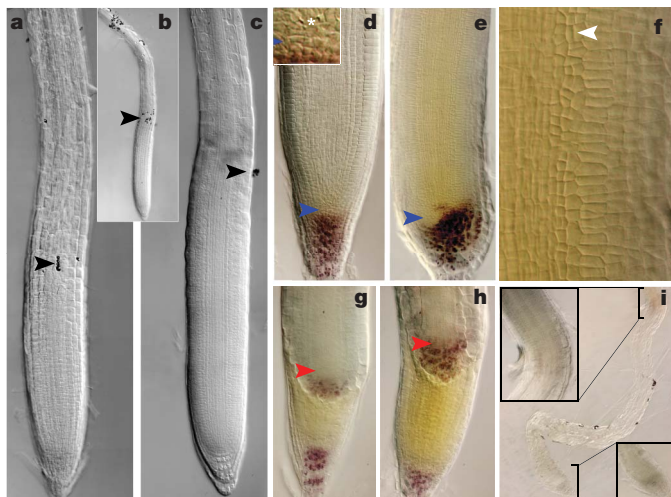
Figure 4 | Inducible expansion of meristem and stem cell area with PLT2–GR fusions. a–c, *35S-PLT2–GR* 7 d.p.g. without dex (**a**) and 1 d after 5 μM dex application (**b, c**). Overview shows positioning of ink toner particles that mark the meristem boundary (black arrowhead) and upper elongation zone boundary at the onset of induction (**b**); the elongation zone boundary is defined as the position where cortical cells rapidly expand. Induced PLT2–GR roots reveal cell division below the meristem boundary and incomplete cell elongation (**c**). **d–f,** *35S-PLT2–GR;pRCH1-RBR* RNAi plants: 10 d.p.g. without dex revealing the two *RBRi*-induced stem cell layers below the quiescent centre (blue arrowhead, inset), asterisk indicates the quiescent centre (**d**); with 3 d of dex application, revealing excessive root cap stem cells (blue arrowhead) and periclinal divisions in the proximal meristem (**e**); magnification with ectopic periclinal divisions (**f**, white arrowhead). **g–i,** Duplication of the stem cell area (red arrowheads) and distal cell types (brown starch granules) in ~10% of 8 d.p.g. *35S-PLT2–GR, pRCH1-RBRi* plants after dex application. Early (**g**), mid- (**h**) and late (**i**) stages of ectopic stem cell centre; note the prolonged activity of both stem cell centres (**i**, inset).

This effect is normally limited by RBR. Low RBR levels in the *RCH1-RBRi* transgenic display limited expansion of the stem cell domain[16] because the PLT levels dictated by the gradient are limiting.

Our data indicate that PLT protein gradients define three outputs in the growing root primordium: stem cell programming, mitotic activity and exit to differentiation. Analysis of PLT target genes will be required to assess how much of the response to graded activity is due to additive concentration effects on the same targets and to differences in target specificity.

Although the molecular link between auxin action and *PLT* gene activation may not be direct[3], auxin distribution and response systems are essential for correct *PLT* gene transcription. This raises the possibility that PLT proteins promote stem cells and transit-amplifying cells as a graded read-out of auxin distribution. In an accompanying paper, we provide evidence that PIN-mediated polar auxin transport establishes a dynamic gradient spanning the root meristem[17]. Hence it is tempting to speculate that an auxin gradient underlies the observed PLT gradients. Classical morphogen systems were conceptualized as independent from the response system. However, several gradients in animal development involve complicated dynamics (for example, ref. 18) and the static concept of positional information is being challenged[19]. We show that PIN polar auxin transport facilitator expression that is essential for correct auxin distribution is regulated by PLT activity, which is a clear example of entanglement between positional information and its response system.

METHODS SUMMARY

Plant work. *plt1-4* and *plt2-2* alleles were described in ref. 3, *plt3-1*, and *bbm-1* and *bbm-2* are salk T-DNA insertion lines 127417, 097021 and 067917, respectively, provided by the Signal Insertion Mutant Library (http://signal.salk.edu/).

The T-DNA insertion in *PLT3* was confirmed by genotyping. The *plt1;plt2;plt3* triple mutant was generated by crossing *plt3-1* to *plt1-4;plt2-2*. *bbm-1* and *bbm-2* were crossed to *plt1-4;plt2-2* and *plt3-1* and allelic combinations were selected from F2 populations. The T-DNA insertion site on *bbm-1* and *bbm-2* lines was verified by genotyping. Primers for genotyping are indicated in Supplementary Table 3. Promoter and genomic sequences were amplified from Col-0 genomic DNA using the primer combinations listed in Supplementary Table 3.. Promoter fragments were fused to the endoplasmic reticulum targeted CFP coding sequence in a pGreenII vector[20]. For translational fusions, *PLT* genomic sequences were fused at the 3′ end to either the *YFP* coding sequence or the carboxy-terminal-encoding region of the rat glucocorticoid (GR) receptor[21] and placed under the control of particular promoters (amplified regions are described in Supplementary Table 3). Promoter swaps were performed by fusing 5.8 kb of *PLT2* and 4.2 kb of *BBM* promoter fragments to the *YFP*-fused *PLT* genomic sequences. Transgenic plants were generated by transforming Col-0 wild-type or *plt1-4;plt2-2* plants, as described[22].

Phenotype analysis and microscopy. Light microscopy[23], confocal microscopy and aniline blue staining[24] of mature embryos was performed as described. Root length was measured, as before[3]. Meristem cell length was measured using ImageJ (v.1.36) and mature cortical cell length as well as fluorescence levels were determined using Zeiss LSM Pascal (3.2SP2) software.

In situ hybridization. Whole-mount RNA *in situ* hybridization was performed as described[11]. The *PLT3* and *BBM* riboprobes, specific for non-conserved sequences downstream of the AP2 repeats, were prepared from templates amplified from complementary DNA (for primers, see Supplementary Table 3). The *PLT1* probe is as in ref. 3; the *PIN1*, *PIN2* and *PIN3* probes are as in ref. 25.

1. Tabata, T. & Takei, Y. Morphogens, their identification and regulation. *Development* **131**, 703–712 (2004).
2. Gurdon, J. B. & Bourillot, P. Y. Morphogen gradient interpretation. *Nature* **413**, 797–803 (2001).
3. Aida, M. *et al.* The PLETHORA genes mediate patterning of the *Arabidopsis* root stem cell niche. *Cell* **119**, 109–120 (2004).
4. Skoog, F. & Miller, C. O. Chemical regulation of growth and organ formation in plant tissues cultured *in vitro. Symp. Soc. Exp. Biol.* **54**, 118–130 (1957).
5. Weigel, D. & Jurgens, G. Stem cells that make stems. *Nature* **415**, 751–754 (2002).
6. Spradling, A., Drummond-Barbosa, D. & Kai, T. Stem cells find their niche. *Nature* **414**, 98–104 (2001).
7. Xu, J. *et al.* A molecular framework for plant regeneration. *Science* **311**, 385–388 (2006).
8. Sabatini, S. *et al.* An auxin-dependent distal organizer of pattern and polarity in the *Arabidopsis* root. *Cell* **99**, 463–472 (1999).
9. Nole-Wilson, S., Tranby, T. L. & Krizek, B. A. AINTEGUMENTA-like (*AIL*) genes are expressed in young tissues and may specify meristematic or division-competent states. *Plant Mol. Biol.* **57**, 613–628 (2005).
10. Birnbaum, K. *et al.* A gene expression map of the *Arabidopsis* root. *Science* **302**, 1956–1960 (2003).
11. Blilou, I. *et al.* The PIN auxin efflux facilitator network controls growth and patterning in *Arabidopsis* roots. *Nature* **433**, 39–44 (2005).
12. Hardtke, C. S. & Berleth, T. The *Arabidopsis* gene *MONOPTEROS* encodes a transcription factor mediating embryo axis formation and vascular development. *EMBO J.* **17**, 1405–1411 (1998).
13. Hellmann, H. *et al. Arabidopsis AXR6* encodes CUL1 implicating SCF E3 ligases in auxin regulation of embryogenesis. *EMBO J.* **22**, 3314–3325 (2003).
14. Dharmasiri, N. *et al.* Plant development is regulated by a family of auxin receptor F box proteins. *Dev. Cell* **9**, 109–119 (2005).
15. Srinivasan, C. *et al.* Heterologous expression of the BABY BOOM AP2/ERF transcription factor enhances the regeneration capacity of tobacco (*Nicotiana tabacum* L.). *Planta* **225**, 341–351 (2007).
16. Wildwater, M. *et al.* The *RETINOBLASTOMA-RELATED* gene regulates stem cell maintenance in *Arabidopsis* roots. *Cell* **123**, 1337–1349 (2005).
17. Grieneisen, V. A., Xu, J., Marée, A. F. M., Hogeweg, P. & Scheres, B. Auxin transport is sufficient to generate a maximum and gradient guiding root growth. *Nature* doi:10.1038/nature06215 (this issue).
18. O'Connor, M. B., Umulis, D., Othmer, H. G. & Blair, S. S. Shaping BMP morphogen gradients in the *Drosophila* embryo and pupal wing. *Development* **133**, 183–193 (2006).
19. Jaeger, J. & Reinitz, J. On the dynamic nature of positional information. *Bioessays* **28**, 1102–1111 (2006).
20. Hellens, R. P., Edwards, E. A., Leyland, N. R., Bean, S. & Mullineaux, P. M. pGreen: a versatile and flexible binary Ti vector for *Agrobacterium*-mediated plant transformation. *Plant Mol. Biol.* **42**, 819–832 (2000).
21. Aoyama, T. & Chua, N. H. A glucocorticoid-mediated transcriptional induction system in transgenic plants. *Plant J.* **11**, 605–612 (1997).

22. Clough, S. J. & Bent, A. F. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis* thaliana. *Plant J.* **16**, 735–743 (1998).

23. Willemsen, V., Wolkenfelt, H., deVries, G., Weisbeek, P. & Scheres, B. The *HOBBIT* gene is required for formation of the root meristem in the *Arabidopsis* embryo. *Development* **125**, 521–531 (1998).

24. Bougourd, S., Marrison, J. & Haseloff, J. Technical advance: an aniline blue staining procedure for confocal microscopy and 3D imaging of normal and perturbed cellular phenotypes in mature *Arabidopsis* embryos. *Plant J.* **24**, 543–550 (2000).

25. Friml, J. *et al.* AtPIN4 mediates sink-driven auxin gradients and root patterning in *Arabidopsis. Cell* **108**, 661–673 (2002).

LETTERS

Purine-mediated signalling triggers eye development

Karine Massé[1], Surinder Bhamra[1], Robert Eason[1], Nicholas Dale[1]* & Elizabeth A. Jones[1]*

A conserved network of eye field transcription factors (EFTFs) underlies the development of the eye in vertebrates and invertebrates[1]. To direct eye development, *Pax6*, a key gene in this network[2,3], interacts with genes encoding other EFTFs such as *Rx1* and *Six3* (refs 4–6). However, the mechanisms that control expression of the EFTFs remain unclear[7]. Here we show that purine-mediated signalling triggers both EFTF expression and eye development in *Xenopus laevis*. Overexpression of ectonucleoside triphosphate diphosphohydrolase 2 (E-NTPDase2)[8], an ectoenzyme that converts ATP to ADP[9], caused ectopic eye-like structures, with occasional complete duplication of the eye, and increased expression of *Pax6*, *Rx1* and *Six3*. In contrast, downregulation of endogenous E-NTPDase2 decreased *Rx1* and *Pax6* expression. E-NTPDase2 therefore acts upstream of these EFTFs. To test whether ADP (the product of E-NTPDase2) might act to trigger eye development through P2Y1 receptors, selective in *Xenopus* for ADP[10,11], we simultaneously knocked down expression of the genes encoding E-NTPDase2 and the P2Y1 receptor. This could prevent the expression of *Rx1* and *Pax6* and eye formation completely. We next measured ATP release[12–14] in the presumptive eye field, demonstrating a transient release of ATP at a time that could plausibly trigger (once converted to ADP) expression of the EFTFs. This surprising role for transient purine-mediated signalling in eye development may be widely conserved, because alterations to the locus of E-NTPDase2 on human chromosome 9 cause severe head and eye defects, including microphthalmia[15–18]. Our results suggest a new mechanism for the initiation of eye development.

To assess the developmental functions of the E-NTPDases, we simultaneously injected the mRNA for the closely related E-NTPDases1–3 (ref. 8) with lineage tracer into a dorsal animal blastomere at the eight-cell stage to target overexpression of this gene to one side of the nervous system. Overexpression of E-NTPDase2 affected eye development in 27 of 41 embryos, causing in some cases complete duplication of the eye on the injected side (Fig. 1A, a). In contrast, overexpression of E-NTPDase1 decreased eye size (11 of 44 embryos; Fig. 1A, b), whereas overexpression of E-NTPDase3 gave a weaker phenotype somewhat similar to that of E-NTPDase2 (4 of 42 embryos, Fig. 1A, c, Supplementary Tables 1a and 2). E-NTPDases differ in their catalytic activity. Like their mammalian orthologues, E-NTPDase1 can metabolize ATP and ADP with roughly equal efficacy, E-NTPDase2 is highly selective for ATP and hardly metabolizes ADP, and E-NTPDase3 has intermediate selectivity for ATP and ADP (Supplementary Fig. 2). The phenotypes elicited by overexpression of these membrane-bound E-NTPDases correlate with their capacity to metabolize ADP.

The eye phenotypes caused by overexpression of E-NTPDase2 (Supplementary Table 1b) included the following: disrupted eye development (Fig. 1A, d); ectopic retinal pigment epithelium (RPE) (Fig. 1A, e); RPE extensions (Fig. 1A, f); and ectopic RPE with an apparent lens (Fig. 1A, g). The antibodies XAR-1, a monoclonal
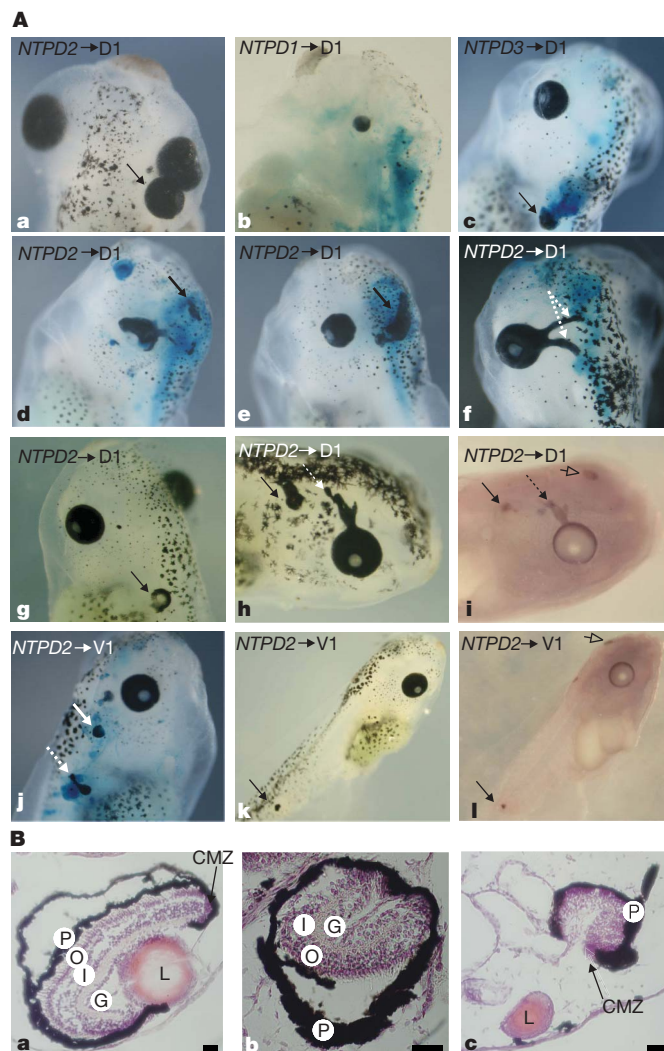

Figure 1 | E-NTPDase2 induces formation of ectopic eye-like structures. **A**, Overexpression of *E-NTPDase* genes causes multiple eye phenotypes. D1 (**a**–**i**) or V1 blastomeres (**j**–**l**) were injected with *Xenopus E-NTPDase (NTPD) 1, 2* or *3* mRNAs. *E-NTPDase2* caused ectopic eyes (black arrows in **a**, **d** and **e**), occasionally with a lens (**g**), abnormal eyes (**d**) or RPE extensions (white dotted arrows, **f**). Pigmented retina (black arrows) and pineal eye (open arrow) (compare **i** with **h**) were immunolabelled with anti-RPE XAR-1. Ventral injection of *E-NTPDase2* gave similar ectopic RPE (arrow) (**j**), sometimes far from the head (**k**, **l**). *E-NTPDase1* caused a small-eye phenotype (**b**). *E-NTPDase3* caused a mild ectopic-eye phenotype (arrow, **c**). **B**, The ectopic eyes exhibit a layered cellular structure. Ectopic eyes (**b** and **c**) of stage-48 embryos show typical organization of retinal layers as in normal eyes (**a**), associated with lenses in some cases (**c**). Scale bars, 30 μm. CMZ, ciliary margin zone; G, ganglion cell layer; I, inner nuclear layer; L, lens; O, outer nuclear layer; P, retinal pigmented layer.

[1]Department of Biological Sciences, Warwick University, Coventry CV4 7AL, UK.
*These authors contributed equally to this work.

anti-RPE (compare Fig. 1A, i with Fig. 1A, h), rhodopsin and the neural marker 2G9 stained the ectopic structures (Supplementary Fig. 3), showing that overexpression of E-NTPDase2 caused the formation of ectopic retinal tissue. Histological sections through the ectopic eyes (Fig. 1B, b and c) showed a layered structure very similar to normal eyes (Fig. 1B, a). This range of eye phenotypes bears a striking resemblance to that reported after overexpression of the paired transcription factor, Pax6 (ref. 19).

Embryos injected in a ventral animal blastomere, fated to give non-neural ectoderm, exhibited ectopic RPE development (8 of 18 embryos; Supplementary Table 1b), usually close to the otic vesicle, but also in the abdomen and along the tail (Fig. 1A, j and k), also positively identified by staining with XAR-1 (Fig. 1A, l). Over-expression of E-NTPDase2 can therefore recruit a cascade of gene expression sufficient to enable eye formation well away from the head structures and the endogenous eye field.

The genes *Rx1* (ref. 5), *Pax6* (refs 2, 3) and *Six3* (ref. 6) are key members of the EFTF network[4]. To determine whether the actions of E-NTPDase2 lie upstream or downstream of the transcription factors encoded by these genes, we analysed their expression by *in situ* hybridization in embryos overexpressing E-NTPDase2 and/or green fluorescent protein (GFP) at stage 13, just after the specification of the eye field[20]. Abnormal expression of *Rx1*, *Pax6* and *Six3* was observed with ectopic patches of staining on the injected side in 21 of 34, 13 of 28, and 14 of 25 embryos, respectively (Fig. 2A, a–c). *Rx1*, *Pax6* and *Six3* expression was not perturbed on the uninjected side or in control GFP-injected embryos (Supplementary Table 3).

Otx2 is a homeodomain transcription factor involved during eye development[21]. The expression of *Otx2* is downregulated at stage 12.5, to reveal an expression pattern that delineates the eye field[4]. E-NTPDase2 overexpression altered *Otx2* expression on the injected side and gave an expansion of the eye field (11 of 23 embryos; Fig. 2A, d). A similar effect was observed with the neural marker *Sox3* (ref. 22) (13 of 27 embryos; Fig. 2A, e). Analysis at stage 23 confirms that ectopic expression for *Otx2* and *Sox3* was observed on the injected side, as was also seen for *Pax6* (Supplementary Fig. 4), indicating an effect on formation of the nervous system. This effect was further confirmed by analysis of embryos at stage 45, which showed duplication of the pineal eye after overexpression of E-NTPDase2 in a small number of cases (Supplementary Fig. 4).

To test whether endogenous E-NTPDase2 is necessary for *Pax6* expression, we knocked down its expression with a specific antisense morpholino oligonucleotide (MO) (Supplementary Fig. 5). Injection of a random control MO (CMO) had no effect on *Pax6* expression at stage 13 (Fig. 2B, a). However, injection of *E-NTPDase2* MO greatly reduced *Pax6* expression in 45 of 48 embryos ($P < 0.0001$ versus CMO; Fig. 2B, b, C). This effect of the *E-NTPDase2* MO was partly rescued by the simultaneous injection of mouse E-NTPDase2 mRNA ($P = 0.016$; Fig. 2B, d, C), the translation of which was unaffected by the *Xenopus*-specific E-NTPDase2 MO (Supplementary Fig. 5a, b) and which when injected by itself enhanced *Pax6* expression (Fig. 2B, c) and caused ectopic eye phenotype (Supplementary Fig. 6 and Supplementary Table 1c).

Overexpression of E-NTPDase1 had no consistent effect on *Pax6* expression (Fig. 2B, e, and Supplementary Table 4). However, if *E-NTPDase1* mRNA was simultaneously injected with *E-NTPDase2* MO, *Pax6* expression was in some cases almost abolished (Fig. 2B, f, C). Overexpression of E-NTPDase1, which can efficiently metabolize ATP and ADP to AMP, will probably exacerbate the effect of knockdown of E-NTPDase2 activity and consequently the decreased production of extracellular ADP. Overexpression of E-NTPDase3 enhanced *Pax6* expression (Fig. 2B, g, and Supplementary Table 4) and partly rescued the effects of the *E-NTPDase2* MO ($P = 0.039$; Fig. 2B, h, C).

E-NTPDase2 seems to be involved in the earliest phase of eye development and acts upstream of the EFTFs encoded by *Pax6*, *Rx1* and *Six3*. As overexpression of E-NTPDases1–3 yields a range

of eye phenotypes that correlates with their enzymatic activity towards ADP, we proposed that signalling by ADP could be critical in activating expression of the EFTFs. *Xenopus* possesses a P2Y1 receptor that is selective for ADP[10,11]. We therefore directly manipulated the expression levels of the P2Y1 receptor with a specific MO (Supplementary Fig. 7), either on its own or in combination with E-NTPDase2.

Knockdown of E-NTPDase2 and the P2Y1 receptor had synergistic effects on EFTF expression at stage 13 (Fig. 3A, B). The *P2Y1* MO alone had no significant effect on *Pax6* and *Rx1* expression (Fig. 3A, c and g). The *E-NTPDase2* MO very significantly decreased *Pax6* and *Rx1* expression ($P = 0.012$ and $P = 0.04$ versus CMO, respectively;
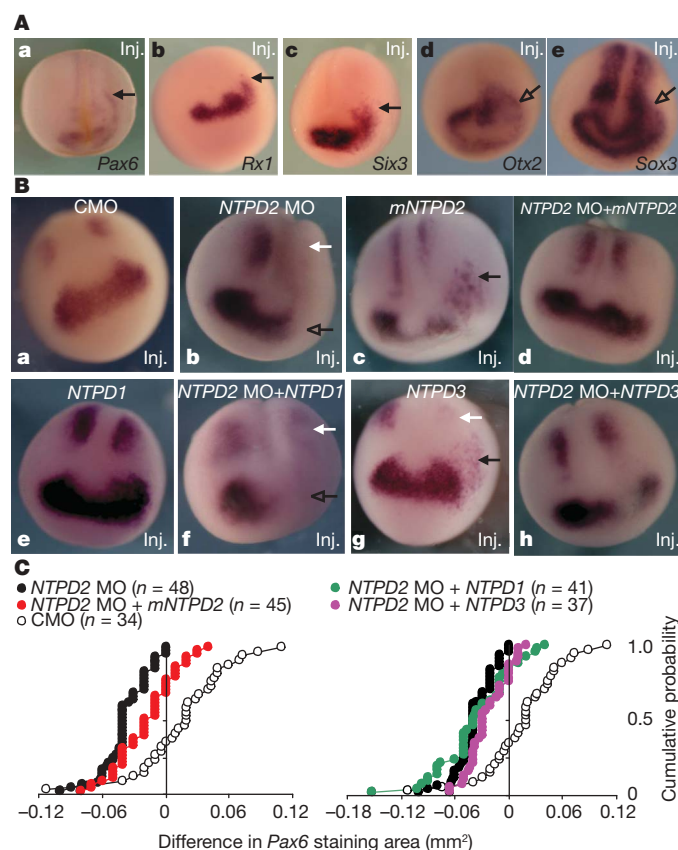

Figure 2 | Mis-expression of *E-NTPDase2* alters EFTF expression.
A, Embryos injected in the D1 blastomere with *E-NTPDase2* mRNA were cultured to stage 13. Ectopic expression (arrow) of the EFTF-encoding genes *Pax6*, *Rx1* and *Six3* (**a–c**) occurred on the injected side (Inj.). The eye field was enlarged on the injected side, as also seen with *Otx2* and *Sox3* expression (open arrow in **d** and **e**). **B**, Embryos injected in the D1 blastomere with mRNAs and MOs were cultured to stage 13. Knockdown of E-NTPDase2 (*NTPD2* MO) decreased *Pax6* expression on the injected side (inj.) in both anterior (open arrow) and posterior (white arrow) domains (**b**). Control MO (CMO) had no effect on *Pax6* expression (**a**). Mouse E-NTPDase2 (*mNTPD2*) caused ectopic expression of *Pax6* (**c**, black arrow). *E-NTPDase2* MO simultaneously injected with *mNTPD2* mRNA rescued *Pax6* expression (compare **d** with **b**). *E-NTPDase1* (*NTPD1*) alone had no significant effect (**e**), whereas *E-NTPDase3* (*NTPD3*) alone caused loss of the posterior *Pax6* expression (white arrow) and ectopic *Pax6* (black arrow) (**g**). *E-NTPDase3*, but not *E-NTPDase1*, was able to rescue the *E-NTPDase2* MO phenotype (compare **h** and **f** with **b**). **C**, Cumulative probability distributions for the effects of control, *E-NTPDase2* MOs, alone or simultaneously injected with *E-NTPDase* mRNAs on the anterior domain of *Pax6* expression. Left: injection of *E-NTPDase2* MO substantially decreased the area of *Pax6* expression (leftward shift). Simultaneous injection of *mE-NTPDase2* mRNA partly rescued this phenotype (shift back to right). Right: simultaneous injection of *E-NTPDase3* mRNA slightly lessened the phenotype caused by *E-NTPDase2* MO, whereas simultaneous injection of *E-NTPDase1* mRNA had little effect.

Fig. 3A, b and f). Simultaneous injection of both MOs almost eliminated *Pax6* and *Rx1* expression ($P < 0.0001$ and $P = 0.001$, respectively; Fig. 3A, a and e, B). Injection of the *E-NTPDase2* and *P2Y1* MOs, either singly or in combination, had no significant effect on the *Sox3* expression that delimits the eye field; however, they did reduce the anterior neural stripe (Fig. 3A, i and j). The CMO had no effect on any of the EFTFs (Fig. 3A, d, h and l).

We next examined the effects of altering *P2Y1* and *E-NTPDase2* expression on eye formation. Overexpression of the P2Y1 receptor gave a weak phenotype (Fig. 4A, a) in comparison with overexpression of E-NTPDase2 (Fig. 4A, b). However, when both the P2Y1 receptor and E-NTPDase2 were overexpressed, the resulting ectopic eye phenotype was greatly exaggerated (Fig. 4A, c). Morpholino knockdown of *P2Y1* expression decreased eye size in comparison with control in about 60% of embryos ($P = 0.057$ versus CMO; Fig. 4B, a, D). Knockdown of E-NTPDase2 gave a similar phenotype in a similar proportion of embryos $P = 0.048$ versus CMO; (Fig. 4B, b, D). However, simultaneous injection of both *P2Y1* and *E-NTPDase2* MO resulted in a markedly enhanced phenotype at stage 45, characterized by either a very small eye or a complete absence of the eye on the injected side ($P = 0.001$ versus CMO; Fig. 4B, c, C, D).

Purinergic signalling may thus lie upstream of the EFTFs and be responsible for triggering the expression of these factors. It should therefore be possible to rescue the effects of knockdown of
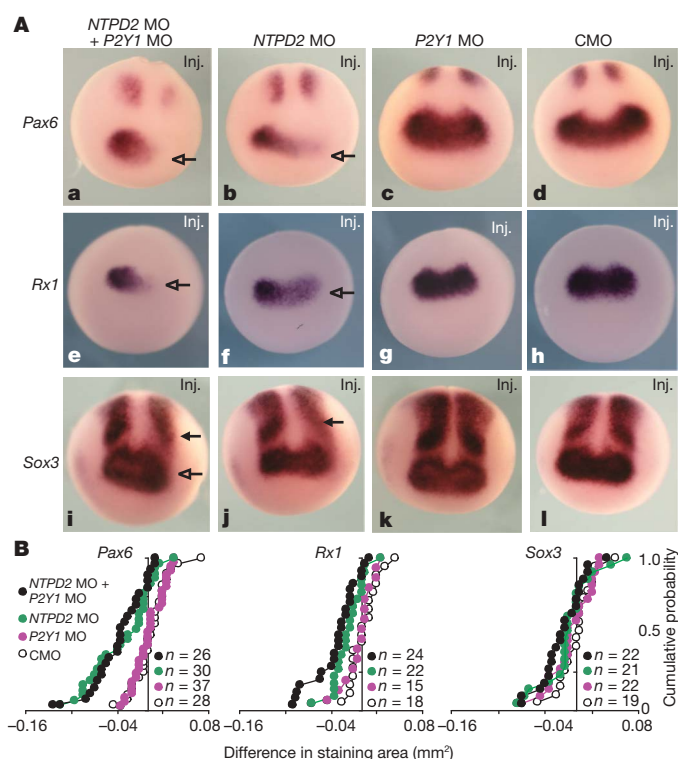
E-NTPDase2 and P2Y1 receptor expression by overexpressing Pax6. Injection of the *E-NTPDase2* and *P2Y1* MOs resulted in a substantial decrease in eye size ($P < 0.0001$ versus CMO; Fig. 5a). Injection of *Pax6* mRNA by itself also gave a typical eye phenotype, as reported by others[19] ($P < 0.0001$ versus CMO; Fig. 5c). However, when the two MOs and *Pax6* mRNA were injected simultaneously, the morphology of the eye on the injected side could be remarkably normal, a phenotype significantly different from those of the double MO phenotype and the control, indicating partial rescue ($P = 0.041$ and $P = 0.028$, respectively; Fig. 5b and Supplementary Fig. 8).

Purinergic signalling by ADP, generated through the actions of E-NTPDase2 and acting through P2Y1 receptors, seems to activate, either directly or indirectly, Pax6 and thus eye formation. Accordingly, the endogenous expression of *E-NTPDase2* and *P2Y1* genes should precede or coincide with the expression of *Pax6* in the eye field. We therefore investigated, by reverse-transcriptase-mediated polymerase chain reaction (RT–PCR), the early expression of these genes in the anterior and posterior neural plate and in the ectodermal or mesodermal/endodermal tissue layers. Expression of *E-NTPDase2*
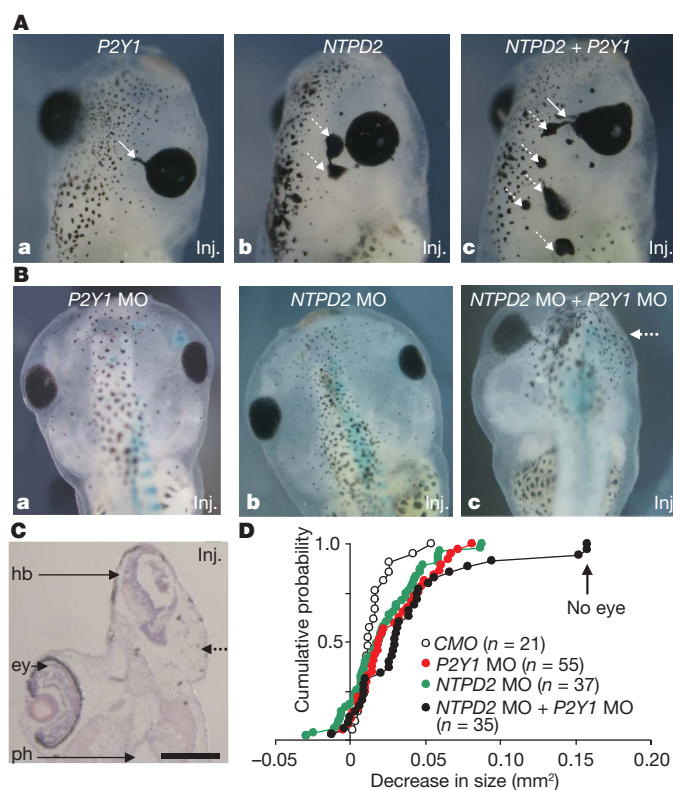

Figure 3 | Both E-NTPDase2 and P2Y1 receptors are necessary for EFTF expression. A, MOs were injected into the D1 blastomere. Knockdown of E-NTPDase2 (*NTPD2* MO) decreased *Pax6* (**b**) and *Rx1* (**f**) staining on the injected side (Inj., open arrow). Knockdown of both E-NTPDase2 and P2Y1 induced a stronger phenotype (**a, e**). Microinjection of both MOs or *E-NTPDase2* MO alone resulted in the loss of a stripe of *Sox3* expression (**i, j**) (filled arrow) but there was little effect on the eye field (open arrow). Microinjection of *P2Y1* MO alone (**c, g, k**) or CMO had no effect on the expression of either marker (**d, h, l**). **B**, Cumulative probability distributions for the effects of control, *P2Y1*, *E-NTPDase2* and combined MOs on *Pax6*, *Rx1* and *Sox3* expression. Injection of both *E-NTPDase2* MO and *P2Y1* MO resulted in a decrease in the area of *Pax6* (anterior domain) and *Rx1* expression on the injected side (leftward shift). This was greater than the phenotype seen with *E-NTPDase2* MO alone. *P2Y1* MO alone had no significant effect on *Pax6* and *Rx1* expression. There was no significant effect of any of the MOs, either alone or in combination, on *Sox3* expression.

Figure 4 | E-NTPDase2 and P2Y1 receptors synergistically affect eye development. A, D1 blastomere-injected *E-NTPDase2* (*NTPD2*) caused the formation of ectopic eyes (**b**, dotted arrow). On rare occasions, *P2Y1* exhibited RPE extensions (**a**, arrow). Simultaneous injection of *E-NTPDase2* with *P2Y1* caused multiple larger ectopic eye-like structures (dotted arrow) coupled with RPE extension (arrow) from the endogenous eye (**c**). **B**, Combined *E-NTPDase2* MO and *P2Y1* MO knockdown ablates eye structures. D1 blastomere knockdown of E-NTPDase2 (*NTPD2* MO) caused a smaller but fully normal eye (**b**) on the injected side. A small effect on phenotype was also seen, by measurement, with *P2Y1* MO alone (**a**). Simultaneous knockdown of E-NTPDase2 and P2Y1 could prevent eye development (**c**, dotted arrow). **C**, Transverse section of the embryo shown in Fig. 3B, c showing a complete loss of eye (dotted arrow) on the injected side. Scale bars, 30 μm. ey, eye; hb, hindbrain; ph, pharynx. **D**, Cumulative probability distributions for the effects of control MO (CMO), *P2Y1* MO, *E-NTPDase2* MO and combined MOs on eye development measured at stage 45. Comparison with the CMO shows that about 30% of the injected embryos showed no eye phenotype. In the remainder, injection with *P2Y1* MO and *E-NTPDase2* MO together resulted in much smaller eyes than injection with CMO.
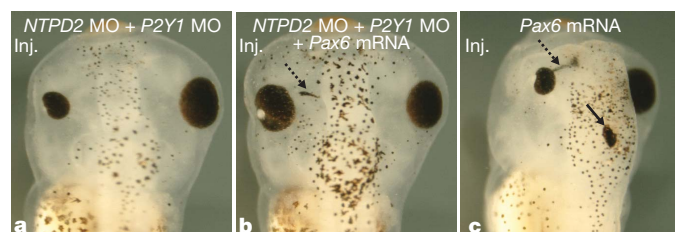
Figure 5 | Pax6 overexpression rescues loss of eye phenotype induced by knockdown of E-NTPDase2 and P2Y1. The small-eye phenotype at stage 45 induced by simultaneous knockdown of E-NTPDase2 and P2Y1 receptor (*NTPDase2* MO + *P2Y1* MO injection through blastomere D1) (**a**) can be rescued by the simultaneous injection of *Pax6* mRNA (**b**). Pax6 overexpression alone (through D1 blastomere) caused ectopic RPE structures (solid arrow) and RPE extensions (**c**)[19] also occasionally seen in **b** (dotted arrows).

becomes detectable in the anterior neural plate at stage 10.5 and thus precedes specification of the eye field[20], characterized by the upregulation of expression of *Pax6* (ref. 4) (Fig. 6a). Furthermore, expression of *E-NTPDase2* is predominantly in the mesodermal/endodermal layer (Fig. 6b). The *P2Y1* gene is expressed preferentially in the anterior neural plate from stage 12.5 and occurs only in the mesoderm/endoderm (Fig. 6a, b). These data indicate that the mesoderm/endoderm tissue is the source of purinergic signalling.

To investigate whether purinergic signalling could provide both spatial and temporal cues for the activation of *Pax6* expression, we directly measured ATP release in the presumptive eye field by using selective biosensors[12]. A single event was recorded from ATP biosensors inserted into the anterior but not the posterior neural plate
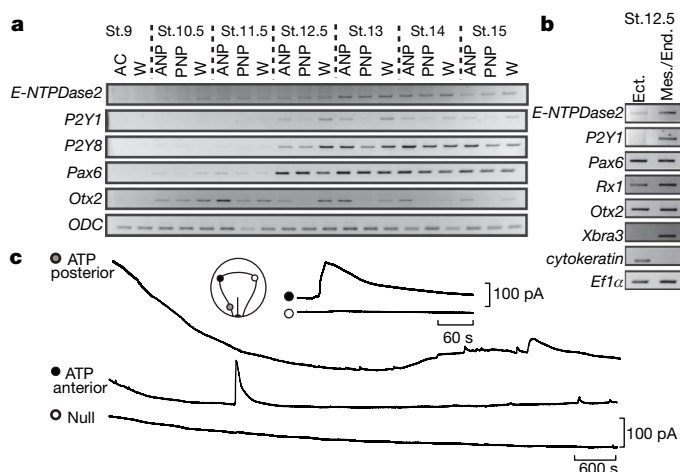
during eye-field development in 10 of 18 embryos (Fig. 6c). No events were recorded on the null biosensor, giving confidence that this signal represented ATP release (Methods). The timing of the event was correlated with the stage of the embryo at the start of the recording. The event had a mean rise time of 25 ± 4.4 s, a duration of 302 ± 53 s and an amplitude of 71 ± 33 pA, equivalent to $1.0 \pm 0.4\,\mu$M ATP. Towards the end of the recording period, ATP release was also observed in the posterior neural plate, indicating that ATP signalling may regulate the development of the posterior neural plate.

We next tested whether pharmacological manipulation of purinergic signalling can either cause or rescue an eye phenotype. Pyridoxal-phosphate-6-azophenyl-2′,4′-disulphonate (PPADS), a P2 antagonist, injected into the blastocoel at $100\,\mu$M, decreased the ectopic expression of *Pax6* caused by the overexpression of E-NTPDase2 (Supplementary Fig. 9a), and MRS2179, a more specific P2Y1 receptor antagonist, when applied similarly decreased the eye phenotype measured at stage 45 (Supplementary Fig. 9b). In the converse experiment, the agonist 2-methylthioadenosine-diphosphate (2MeSADP) was able to rescue the effects of the *E-NTPDase2* MO on *Pax6* expression (Supplementary Fig. 9a). We also applied the P2Y1 agonist 2-MeSADP by means of agarose beads soaked in a 1 mM solution. Beads inserted near the presumptive eye field at stage 12.5 occasionally produced RPE extensions (Supplementary Fig. 9c) typical of overexpression of E-NTPDase2 or Pax6.

Our data provide compelling evidence that a discrete purinergic signalling event localized to the anterior neural plate is critical for triggering the expression of EFTFs and hence initiating eye development (Supplementary Fig. 1). Because mutations to *Pax6* can give rise to severe defects of the nervous system[23–25], we expected similar alterations from manipulation of the purinergic signalling pathway upstream of *Pax6*. Accordingly, alteration of *E-NTPDase2* expression affected both brain development (Supplementary Fig. 4) and *Pax6* expression in the posterior neural plate (Fig. 2B, b)—a region destined to form the nervous system. Purinergic signalling thus orchestrates the development of both the anterior nervous system and the eye. E-NTPDase2 has been implicated in the differentiation of stem cells in the vertebrate nervous system[26,27].

Human eye and brain development may also depend on E-NTPDase2 activity: spontaneous mutations at 9q34, the region bearing the locus of the *E-NTPDase2* gene, consistently give head and brain abnormalities that include a range of eye defects including microphthalmia, analogous to the phenotypes we have reported in the frog[15–18]. Our studies raise several fascinating questions such as the identity of the ATP-releasing cells and the chain of events that connect activation of the P2Y1 receptor in the mesoderm/endoderm to initiation of EFTF expression in the ectoderm.

Figure 6 | All components of purinergic signalling are present for the initiation of eye development. a, *E-NTPDase2* and *P2Y1* are expressed in the eye field. *E-NTPDase2* transcripts were detected from stage (St.) 10.5 in the presumptive anterior neural plate (ANP) and posterior neural plate (PNP), whereas *Pax6* was detected later, from stage 11.5. *E-NTPDase3* was first expressed at stage 13 in the ANP but *E-NTPDase1* was not detected at any stage (data not shown). *P2Y1* (ANP > PNP) and *P2Y8* (ANP = PNP) mRNAs were detected from stage 12.5. *Otx2* demonstrated the accuracy of the dissections, and *ODC* was the loading control. AC, animal cap; W, whole embryo. **b**, *E-NTPDase2* and *P2Y1* are expressed in the mesodermal layer. *E-NTPDase2* was expressed at higher levels in dissected mesodermal/endodermal (Mes./End.) layers than in the ectoderm (Ect.). *P2Y1* was only detected in mesoderm/endoderm dissection. *Otx2*, *Rx1* and *Pax6* were expressed in both fractions. *Xbra3* and *cytokeratin* demonstrated the accuracy of the dissections. *EF1α* was the loading control. **c**, ATP is released in the neural plate during development. The placement of biosensors is indicated in the diagram. A single large transient event was seen on the ATP biosensor inserted into the anterior neural plate. This event had a rapid rise time and a slower decaying phase (inset) and was unaccompanied by any equivalent signal on the null or posterior ATP biosensors.

METHODS SUMMARY

Embryo culture dissection and microinjections. *Xenopus laevis* embryos were staged, cultured and dissected by standard methodology. All microinjections of mRNAs, morpholinos or combinations of both were performed into defined blastomeres at the eight-cell stage together with lineage tracer.

RT–PCR. RNAs were isolated and complementary DNAs were reverse-transcribed by following standard procedures from whole or dissected embryos. Non-radioactive PCRs were performed with gene-specific primers.

Whole-mount *in situ* hybridization, immunohistochemistry and sectioning. Experimentally manipulated embryos were hybridized to digoxigenin-labelled RNA probes produced *in vitro* from cDNA clones encoding eye and brain markers. Colour reactions were performed and the embryos were bleached before scoring. Immunohistochemically stained embryos were analysed by standard protocols and colour reactions.

Embryo measurement and statistical analysis. Areas of positive *in situ* hybridization were measured with image analysis software (LuciaG) and recorded. The dimensions of the eyes were measured along orthogonal axes and the product of these values was used as a proxy for volume. Differences between injected and uninjected sides were ranked and cumulative probability distributions were plotted. Statistical analysis was performed by Kruskall–Wallis analysis of

variance, with pairwise comparisons by the Kolmogorov–Smirnov test (probability values for these comparisons are given in text where appropriate).

Biosensor recordings. ATP biosensors were used to record ATP release from embryos at the mid-gastrula stage. Null biosensors acted as negative controls and ATP sensors were calibrated at the end of each experiment.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Chow, R. & Lang, R. Early eye development in vertebrates. *Annu. Rev. Cell Dev. Biol.* **17,** 255–296 (2001).
2. Hill, R. E. *et al.* Mouse *Small eye* results from mutations in a paired-like homeobox-containing gene. *Nature* **354,** 522–525 (1991).
3. Quiring, R., Walldorf, U., Kloter, U. & Gehring, W. J. Homology of the *eyeless* gene of *Drosophila* to the *Small eye* gene in mice and Aniridia in humans. *Science* **265,** 785–789 (1996).
4. Zuber, M., Gestri, G., Viczian, A., Barsacchi, G. & Harris, W. A. Specification of the vertebrate eye by a network of eye field transcription factors. *Development* **130,** 5155–5167 (2003).
5. Mathers, P., Grinberg, A., Mahon, K. & Jamrich, M. The *Rx* homeobox gene is essential for vertebrate eye development. *Nature* **387,** 603–607 (1997).
6. Loosli, F., Winkler, S. & Wittbrodt, J. Six3 overexpression initiates the formation of ectopic retina. *Genes Dev.* **13,** 649–654 (1999).
7. Esteve, P. & Bovolenta, P. Secreted inducers in vertebrate eye development: more functions for old morphogens. *Curr. Opin. Neurobiol.* **16,** 13–19 (2006).
8. Massé, K., Eason, R., Bhamra, S., Dale, N. & Jones, E. A. Comparative genomic and expression analysis of the conserved *NTPDase* gene family in *Xenopus. Genomics* **87,** 366–381 (2006).
9. Zimmermann, H. Extracellular metabolism of ATP and other nucleotides. *Naunyn-Schmiedebergs Arch. Pharmacol.* **362,** 299–309 (2000).
10. Brown, P. & Dale, N. Modulation of K$^+$ currents in *Xenopus* spinal neurons by p2y receptors: a role for ATP and ADP in motor pattern generation. *J. Physiol. (Lond.)* **540,** 843–850 (2002).
11. Cheng, A. W. *et al.* cDNA encodes *Xenopus* P2Y(1) nucleotide receptor: expression at the neuromuscular junctions. *Neuroreport* **14,** 351–357 (2003).
12. Llaudet, E., Hatz, S., Droniou, M. & Dale, N. Microelectrode biosensor for real-time measurement of ATP in biological tissue. *Anal. Chem.* **77,** 3267–3273 (2005).
13. Pearson, R. A., Dale, N., Llaudet, E. & Mobbs, P. ATP released via gap junction hemichannels from the pigment epithelium regulates neural retinal progenitor proliferation. *Neuron* **46,** 731–744 (2005).
14. Gourine, A. V., Llaudet, E., Dale, N. & Spyer, K. M. ATP is a mediator of chemosensory transduction in the central nervous system. *Nature* **436,** 108–111 (2005).
15. Neas, K. R. *et al.* Three patients with terminal deletions within the subtelomeric region of chromosome 9q. *Am. J. Med. Genet. A* **132,** 425–430 (2005).
16. Allderdice, P. W. *et al.* Duplication 9q34 syndrome. *Am. J. Hum. Genet.* **35,** 1005–1019 (1983).
17. Hampshire, D. J. *et al.* MORM syndrome (mental retardation, truncal obesity, retinal dystrophy and micropenis), a new autosomal recessive disorder, links to 9q34. *Eur. J. Hum. Genet.* **14,** 543–548 (2006).
18. Yatsenko, S. A. *et al.* Deletion 9q34.3 syndrome: genotype–phenotype correlations and an extended deletion in a patient with features of Opitz C trigonocephaly. *J. Med. Genet.* **42,** 328–335 (2005).
19. Chow, R., Altmann, C., Lang, R. & Hemmati-Brivanlou, A. Pax6 induces ectopic eyes in vertebrates. *Development* **126,** 4213–4222 (1999).
20. Li, H., Tierney, C., Wen, L., Wu, J. Y. & Rao, Y. A single morphogenetic field gives rise to two retina primordia under the influence of the prechordal plate. *Development* **124,** 603–615 (1997).
21. Martinez-Morales, J. R., Signore, M., Acampora, D., Simeone, A. & Bovolenta, P. *Otx* genes are required for tissue specification in the developing eye. *Development* **128,** 2019–2030 (2001).
22. Collignon, J. *et al.* A comparison of the properties of *Sox-3* with *Sry* and two related genes, *Sox-1* and *Sox-2. Development* **122,** 509–520 (1996).
23. Glaser, T. *et al.* PAX6 gene dosage effect in a family with congenital cataracts, aniridia, anophthalmia and central nervous system defects. *Nature Genet.* **7,** 463–471 (2004).
24. Stoykova, A., Fritsch, R., Walther, C. & Gruss, P. Forebrain patterning defects in *Small eye* mutant mice. *Development* **122,** 3453–3465 (1996).
25. Mitchell, T. N. *et al.* Polymicrogyria and absence of pineal gland due to *PAX6* mutation. *Ann. Neurol.* **53,** 658–663 (2003).
26. Braun, N. *et al.* Expression of the ecto-ATPase NTPDase2 in the germinal zones of the developing and adult rat brain. *Eur. J. Neurosci.* **17,** 1355–1364 (2003).
27. Mishra, S. K. *et al.* Extracellular nucleotide signaling in adult neural stem cells: synergism with growth factor-mediated cellular proliferation. *Development* **133,** 675–684 (2006).

METHODS

DNA constructs. Full-length *Xenopus E-NTPDase1*, *E-NTPDase2* and *E-NTPDase3* cDNAs were isolated from IMAGE clones[8], cloned into pBF. The IMAGE clones coding for mouse *E-NTPDase2* cDNA (IMAGE clone 3582550, NCBI; accession number BC011241) and *Xenopus P2Y1* cDNA (IMAGE clone 5507066, NCBI; accession number CF284315) were ordered from the UK MRC HGMP Resource Centre, and the full coding region was inserted into pBF.

Embryo culture, dissection and microinjections. Embryos, staged as described[28], were obtained and cultured as described previously. Embryos at different stages were dissected in Barth X to isolate the animal cap (stage 9) and the anterior neural plate from the posterior neural plate (from stage 10.5 to stage 15). Fine dissection of the anterior neural plate of stage 12.5 embryos was performed to isolate the outer ectoderm from the inner mesodermal and endodermal layers. Capped synthetic mRNAs were generated by transcription *in vitro* from the expression clones by using the mMessage mMachine kit (Sp6 RNA polymerase; Ambion) and 5 ng was injected into each embryo. All microinjections were performed into a D1 or V1 blastomere (as indicated in the figure legends) of eight-cell-stage embryos. *E-NTPDase2* (5′-gccatgttgggctcttgagtgcctc-3′) and *P2Y1* (5′-gagagaaagacttctgtcatgatct-3′) MOs were designed and supplied by GeneTools. The MO (5 ng nl^{-1}) was injected alone (20 ng) or in combination with mRNA or with another MO. The random CMO (GeneTools) was used as a control. Either *GFP* (4 ng) or *β-galactosidase* (*LacZ*) (2 ng) mRNA was simultaneously injected as a lineage tracer. X-Gal staining was performed as described[29]. Each experiment was performed on at least two independent batches of embryos; data presented and analysed in the text are from a representative batch.

RT–PCR. Total RNA from whole or dissected embryos was isolated and used for reverse transcription as described[30]. PCR was performed with non-radioactive nucleotides (Supplementary Table 5). For each experiment the quantity of input cDNA was determined by equalization of the samples with a housekeeping gene.

Whole-mount *in situ* hybridization and immunohistochemistry. Microinjected embryos were fixed in MEMFA (0.5 M MOPS pH 7.4, 100 mM EGTA, 1 mM MgSO$_4$, 4% formaldehyde) and hybridized (see ref. 31) with antisense RNA probes produced *in vitro* from cDNA clones. A 800-base-pair fragment from *Sox3* was cloned into pGEMT-easy, transcribed with Sp6 RNA polymerase and linearized with *Sac*II. Clones containing the templates for *Pax6*, *Rx1*, *Six3* and *Otx2* were generated in accordance with published protocols[4,32–34]. The probe was labelled by using a digoxigenin labelling kit (Roche) and the hybridization was revealed with sheep anti-digoxigenin-alkaline phosphatase antibody (Roche) and 4-nitroblue tetrazolium chloride/5-bromo-4-chloro-3-indolylphosphate substrate (NBT/BCIP; Roche). Embryos were bleached by using standard protocols. Whole-mount immunohistochemistry was performed on injected embryos fixed in MEMFA and dehydrated in methanol. Embryos were rehydrated, bleached and immunostained by using standard protocols with the monoclonal anti-RPE antibody XAR-1. The colour reaction was performed with NBT/BCIP.

Embedding and sectioning. Wax sections (12 μm thick) were cut from stage-48 embryos[35] and stained with 20% Harris haematoxylin solution (Sigma) and 1% eosin (BDH Laboratory Supplies). The sections were mounted with DePex (BDH Laboratory Supplies).

Measurement of embryos and statistical analysis. A Nikon SMZ1500 microscope and digital DXM1200F camera were used to photograph the embryos. The midline of each embryo was identified and the anterior domain of *Pax6* staining, *Rx1* domain or the eye field delimited by *Sox3* expression on both injected and uninjected sides was measured with the software LuciaG. The difference in area was ranked and plotted as a cumulative probability distribution. The size of eyes of stage-45 embryos was measured along the anterior–posterior and medio-lateral axes by using LuciaG. The product of these two measurements was calculated (in mm^2) and used as a proxy statistic for eye volume. The decrease in eye size between the injected and uninjected sides was ranked and plotted as a cumulative probability distribution. The *G*-test was used to compare the frequency of the phenotypes evoked by each experimental manipulation (Supplementary Table 2). A combination of the Kruskal–Wallis one-way analysis of variance and pairwise comparisons performed with the Kolmogorov–Smirnov two-sample test was used to analyse the cumulative probability distributions presented in Figs 2–4 and Supplementary Fig. 8 (Supplementary Information). For the analysis in Figs 2 and 3, two-tailed comparisons were used; for Fig. 4 and Supplementary Fig. 8, in which our hypothesis predicts a particular result, one-tailed comparisons were used.

Biosensor recordings. The ATP biosensors have been described elsewhere[12,13]. Those used here were modified with an internal screening layer to enhance the selectivity and greatly reduce their sensitivity to non-specific electroactive reagents. The sensing portion was 500 μm long and 50 μm in diameter. Embryos at late stage 11.5 were placed in a small chamber and wedged between fine tungsten pins so that the developing neural plate was uppermost. The biosensors were inserted into the embryo, to a depth of about 300 μm, through a small hole made in the correct location with a sharpened tungsten pin. Glycerol (5 mM), an essential co-substrate for the biosensor operation, was included in 1/10 Barth X recording medium along with 1 mM MgCl$_2$. The null biosensors, lacking the ATP-detecting enzymes, acted as a control. They were sensitive to any possible interference but did not respond to ATP. A continuous recording was made for at least 2 h (Stage 13). At the end of the experiment, the sensors were withdrawn from the embryo and calibrated with ATP to allow conversion of the sensor signal to units of concentration.

28. Nieuwkoop, P. D. & Faber, J. *Normal Table of* Xenopus laevis *(Daudin)*. (Garland, New York, 1994).
29. Bourguignon, C., Li, J. & Papalopulu, N. XBF-1, a winged helix transcription factor with dual activity, has a role in positioning neurogenesis in *Xenopus* competent ectoderm. *Development* **125**, 4889–4900 (1998).
30. Barnett, M. W., Old, R. W. & Jones, E. A. Neural induction and patterning by fibroblast growth factor, notochord and somite tissue in *Xenopus. Dev. Growth Differ.* **40**, 47–57 (1998).
31. Harland, R. *In situ* hybridisation: an improved wholemount method for *Xenopus* embryos. *Methods Cell Biol.* **36**, 685–695 (1991).
32. Hirsch, N. & Harris, W. A. *Xenopus* Pax-6 and retinal development. *J. Neurobiol.* **32**, 45–61 (1997).
33. Casarosa, S., Andreazzoli, M., Simeone, A. & Barsacchi, G. *Xrx1*, a novel *Xenopus* homeobox gene expressed during eye and pineal gland development. *Mech. Dev.* **61**, 187–198 (1997).
34. Pannese, M. *et al.* The *Xenopus* homologue of *Otx2* is a maternal homeobox gene that demarcates and specifies anterior body regions. *Development* **12**, 707–720 (1995).
35. Haldin, C. E., Nijjar, S., Massé, K., Barnett, M. W. & Jones, E. A. Isolation and growth factor inducibility of the *Xenopus laevis* Lmx1b gene. *Int. J. Dev. Biol.* **47**, 253–262 (2003).

LETTERS

Functional architecture of the retromer cargo-recognition complex

Aitor Hierro[1]*, Adriana L. Rojas[1]*, Raul Rojas[2], Namita Murthy[2], Grégory Effantin[3], Andrey V. Kajava[4], Alasdair C. Steven[3], Juan S. Bonifacino[2] & James H. Hurley[1]

The retromer complex[1,2] is required for the sorting of acid hydrolases to lysosomes[3–7], transcytosis of the polymeric immunoglobulin receptor[8], Wnt gradient formation[9,10], iron transporter recycling[11] and processing of the amyloid precursor protein[12]. Human retromer consists of two smaller complexes: the cargo recognition VPS26–VPS29–VPS35 heterotrimer and a membrane-targeting heterodimer or homodimer of SNX1 and/or SNX2 (ref. 13). Here we report the crystal structure of a VPS29–VPS35 subcomplex showing how the metallophosphoesterase-fold subunit VPS29 (refs 14, 15) acts as a scaffold for the carboxy-terminal half of VPS35. VPS35 forms a horseshoe-shaped, right-handed, α-helical solenoid, the concave face of which completely covers the metal-binding site of VPS29, whereas the convex face exposes a series of hydrophobic interhelical grooves. Electron microscopy shows that the intact VPS26–VPS29–VPS35 complex is a stick-shaped, flexible structure, approximately 21 nm long. A hybrid structural model derived from crystal structures, electron microscopy, interaction studies and bioinformatics shows that the α-solenoid fold extends the full length of VPS35, and that VPS26 is bound at the opposite end from VPS29. This extended structure presents multiple binding sites for the SNX complex and receptor cargo, and appears capable of flexing to conform to curved vesicular membranes.

The retromer cargo-recognition complex consists of the 38-kDa VPS26, 20-kDa VPS29 and 92-kDa VPS35 subunits. The structures of the two smaller subunits have been determined in isolation. VPS26 is a structural cousin of the arrestins[16], a family of trafficking proteins that directly bind to cell surface receptors and direct their internalization. VPS29 has a metallophosphoesterase fold[14,15] that can bind two metal ions. Compared to functional metallophosphoesterases, a key histidine residue that serves as a catalytic base in the metallophosphoesterase active site is replaced by Phe 63. Thus, VPS29 is completely inactive with respect to generic phosphatase substrates[15]. However, metal-dependent activity *in vitro* against a phosphorylated peptide from a major retromer cargo, the cation-independent mannose 6-phosphate receptor (CI-MPR), has been reported[17]. Despite its centrality to multiple trafficking pathways, the precise function of retromer has been enigmatic. Various proposals have emphasized potential roles as a coat, adaptor, or cargo protein phosphatase. Here we take a structural approach to gain insight into the function of retromer.

The crystal structure of a VPS29–VPS35 subcomplex, containing the C-terminal 40% of the large VPS35 subunit, was determined at 2.8 Å resolution (Fig. 1a, Supplementary Fig. 1 and Supplementary Table 1). The C-terminal portion of VPS35 consists of a single

right-handed superhelix with a pitch of 12 Å and a total of 13 helices (Fig. 1a and Supplementary Fig. 2). VPS35 resembles many other helical solenoid proteins, including other important proteins in coated vesicle trafficking. These include the trunk domains of the large subunits of the AP adaptor protein complexes[18,19] (Supplementary Fig. 3), the function of which is to link clathrin to cargo and membranes, the helical repeat regions of the heavy chain of clathrin itself[20,21], and the Sec13–Sec31 assembly unit of the COPII coat[22]. VPS35 wraps itself nearly halfway around the VPS29 subunit, burying 3,300 Å^2 of solvent-accessible surface area (Fig. 1b, c). The VPS35 binding site on VPS29 includes the entire metal-binding site, as well as flanking residues. Ile 91 of VPS29, previously shown to contact VPS35 (ref. 15), interacts extensively with VPS35 near the centre of the VPS29–VPS35 interface (Fig. 1a). VPS35 and its C-terminal fragment are both poorly stable when expressed in isolation, consistent with a fold that depends on scaffolding by VPS29 for its stability. The even-numbered helices of the convex face of VPS35 are splayed apart from each other because the solenoid is curved. This surface thus consists of a series of ridges separating hydrophobic grooves, suggestive of potential cargo-binding sites (Fig. 1d). These grooves are reminiscent of the CI-MPR binding site on the helical solenoidal VHS domains of the GGA clathrin adaptor proteins[23,24]. The binding site for Vps10 to yeast Vps35 (ref. 25) maps to part of this surface (Fig. 1d).

Two models have been proposed for the function of VPS29. In the first, VPS29 is a scaffold for interaction with VPS35 and SNX1 (ref. 15) with no phosphatase activity of functional significance. In the second, VPS29 is an active protein phosphatase for the CI-MPR and perhaps other receptors[17]. The complete burial of the putative VPS29 metallophosphoesterase active site prompted us to revisit the report of CI-MPR dephosphorylation activity by retromer[17]. Working with preparations purified to homogeneity, neither the complete VPS26–VPS29–VPS35 complex nor the VPS29–VPS35 subcomplex had detectable activity (Supplementary Fig. 4). Furthermore, VPS29 alone had no detectable activity.

The lack of *in vitro* phosphatase activity suggests, but does not prove, that the main function for the metallophosphoesterase site is to provide a scaffold for VPS35. VPS29 must associate with VPS35 to maintain its known biological functions in sorting[15]. We considered whether a conformational change in VPS35 could expose the VPS29 metallophosphoesterase site without completely disassembling the complex. One region of VPS35 covers the putative VPS29 active site and is thus potentially an inhibitory contact site, whereas another region is remote from the putative active site and unlikely to be involved in inhibition (Fig. 1b, c). Each of these two

regions comprises roughly half of the total contact area. The scaffold model predicts that mutations that selectively disrupt the putative inhibitory region should prevent assembly of the complex. The phosphatase model, on the other hand, predicts that these mutations should allow integration into the retromer but lead to constitutive exposure of the metallophosphoesterase site. HeLa cells were transfected with mammalian expression plasmids encoding haemagglutinin (HA)-tagged wild-type or mutant VPS35 constructs carrying multiple substitutions on residues (Fig. 1b, grey residues) covering the putative VPS29-phosphatase active site or residues (Fig. 1b, purple residues) distal to this region. After transfection, the proteins were isolated by immunoprecipitation and analysed by SDS–polyacrylamide gel electrophoresis (PAGE) and immunoblotting with antibodies to VPS26 and VPS29 (Fig. 2a, top). We found that wild-type HA–VPS35 can assemble with endogenous VPS29 and VPS26. In contrast, all of the HA-tagged VPS35 mutants in both regions prevented assembly with VPS29, consistent with the scaffolding model. All of the recombinant HA-tagged VPS35 constructs were expressed at similar levels in the transfected cells (Fig. 2a, bottom). In addition, these HA-tagged VPS35 mutants all assembled with VPS26 (Fig. 2a), indicating that they are capable of proper folding and that the VPS26 and VPS29 binding sites are independent of one another. We used double-labelling fluorescent microscopy to study the cellular distribution of these HA-tagged constructs (Supplementary Fig. 5a). Unlike wild-type HA–VPS35, which is mostly endosomal, all of the mutant HA–VPS35 constructs showed a diffuse cytosolic distribution.

Similarly, we transfected wild-type Myc-tagged VPS29 and versions harbouring mutations of VPS35-interacting residues surrounding the metallophosphoesterase binding site (Fig. 1c, grey residues) and remote from it (Fig. 1c, purple residues). These constructs were immunoprecipitated and analysed by SDS–PAGE and immunoblotting using antibodies to VPS35 and VPS26 (Fig. 2b, top). All but one (L101D) of the characterized Myc-tagged VPS29 mutants failed to assemble with endogenous VPS35, re-emphasizing that all regions of the interface are important for assembly. We also tested nine mutants of the VPS29 metallophosphoesterase site and found that only three (D8N, H86A and H117A) assembled partially with VPS35, whereas the remaining mutants did not assemble (Supplementary Table 2). The levels of these mutants were much lower than wild-type Myc-tagged VPS29 and almost undetectable in lysates by immunoblot analysis (Supplementary Table 2). This led us to conclude that an intact metal-binding site is required for the stability of VPS29.

Indeed, the use of a metallophosphoesterase fold for non-catalytic purposes may be common. Of 5,307 putative metallophosphoesterase catalytic domains (see Methods), only 41% of the sequences conserve eight key residues directly involved in metal binding or catalysis. Even though catalytically inactive metallophosphoesterases seem to comprise most of this vast protein fold family, little is known about their functions. It will be interesting to learn whether the VPS29 example is a paradigm for other members of this class. How might so many such folds have evolved? The structure of full-length protein phosphatase 5 shows that its TPR domain, a helical solenoid, is inserted into and sterically blocks the active site of the catalytic domain[26], and suggests the nature of a potential evolutionary precursor. We suggest that most of the metal ligands were retained during evolution for stability reasons, despite their lack of catalytic function.
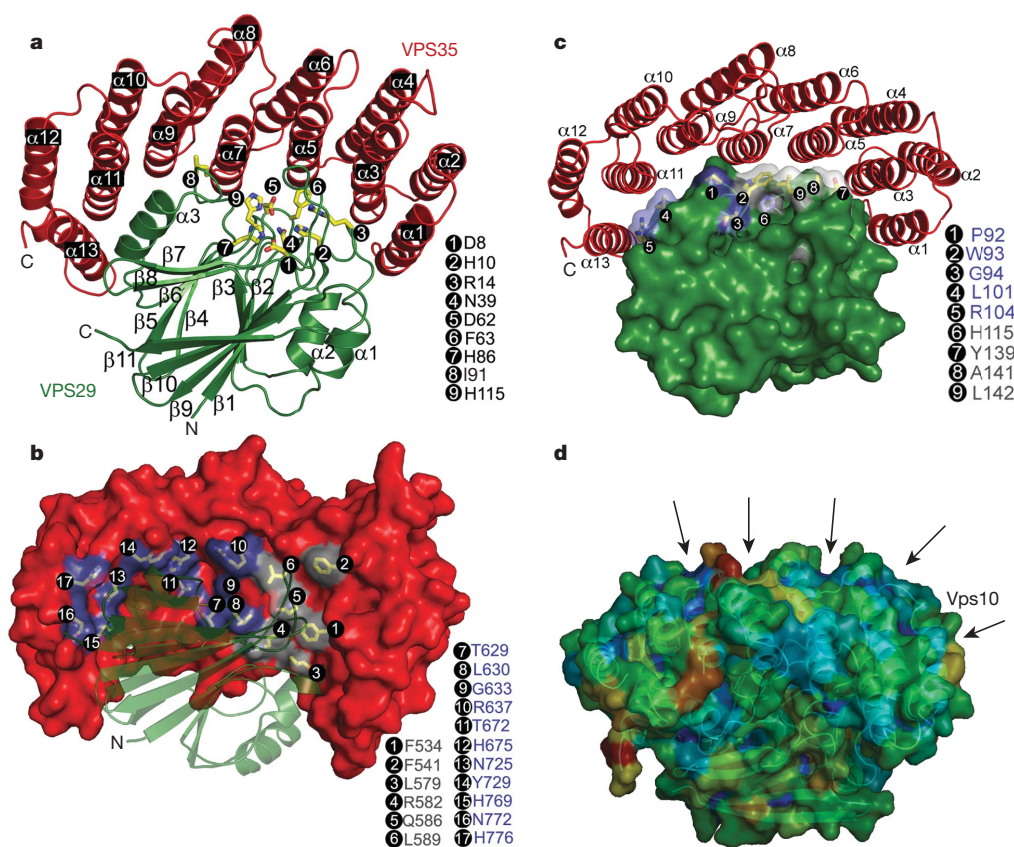

Figure 1 | Structure of the VPS29–VPS35 subcomplex. a, VPS29 is green and VPS35 red. **b**, The surface of VPS35 is shown, with the residues blocking the metallophosphoesterase site of VPS29 in grey, and other residues that contact VPS29 in purple. **c**, The surface of VPS29 is shown, with residues surrounding the metallophosphoesterase site in light blue, and other VPS35-contacting residues in purple. **d**, Hydrophobic grooves on the outer surface of VPS35 are formed between even-numbered helices. The probability of the surface to participate in ligand binding was coloured from lowest to highest in a blue to red gradient using the hotpatch server (http:// hotpatch.mbi.ucla.edu/). Structural figures were generated with pymol (http://www.pymol.org/).

The intact recombinant VPS26–VPS29–VPS35 complex was found to contain an equimolar ratio of the three subunits by Coomassie-stained SDS–PAGE (Fig. 2c) and by quantitative amino acid analysis (data not shown). The elution of the complex on gel filtration chromatography (data not shown) is potentially consistent with either an elongated 1:1:1 complex or a globular 2:2:2 complex. Given that the electron microscopy described below indicates that the structure is elongated and occupies a volume consistent with a 1:1:1 complex, we favour this interpretation. Through co-expression of truncated recombinant proteins, we mapped the interaction of VPS26 and VPS35 to the C-terminal lobe of VPS26 and the N-terminal 165 amino acids[4] of VPS35 (Fig. 2d). The organization of the large subunit, VPS35, is central to understanding the structure of retromer. Analysis of the VPS35 sequence showed that it consists of regular repeats of a two-helix motif (Fig. 3a), strongly suggesting that the α-solenoid visualized in the crystallized fragment extends throughout essentially the entire structure. The total number of helices is anticipated to be 34, including the 13 in VPS35-C. This corresponds to a contour length of ∼204 Å for the complete molecule, given a pitch of 12 Å per two-helix turn of the solenoid in the crystallized portion. The conformations of other α-solenoids vary from ring-like to quasi-linear[27,28]. How is the putative α-solenoid arranged in this case? To investigate this question, we performed negative staining electron microscopy. The images (for example, see Supplementary Fig. 6) exhibit some variability but many molecules are filamentous and about 210 Å long. To enhance detail, we performed image averaging after alignment and classification. The class averages exhibit some variation in local structure but are rather consistent in length, between 210 and 220 Å (Fig. 3b). Each has five or six stain-excluding densities. These data can be accounted for by a model consisting of the VPS29–VPS35-C heterodimer at one end, extended by ∼135 Å of α-solenoid with VPS26 attached near its distal end (Fig. 3c, d). This model was

tested by generating resolution-limited projections from the model for comparison with the negatively stained class averages (Fig. 3c). At one end, the structure is consistent with projections of the VPS29–VPS35-C subcomplex. The respective class averages correspond to different side views of this assembly, with some flexing about its mid-section (Supplementary Fig. 7). These data strongly support the inferred continuation of the VPS35 α-solenoid in an approximately linear conformation, with VPS26 bound near its end.

The overall structural model suggests a mechanism for cooperative binding to multiple membrane-anchored molecules (Fig. 4a). The membrane-binding SNX1 subunit interacts with a groove on VPS29 (at Leu 25 and Leu 152)[15] on the face distal to VPS35, and with two different sites on the N- and C-terminal halves of VPS35 (ref. 4). The N- and C-terminal halves of yeast Vps35 have been shown to interact with the distinct cargoes A-ALP and Vps10, respectively[25]. The Vps35 binding site on the A-ALP cytosolic domain is within 20 residues of the A-ALP transmembrane region[29], dictating that the N-terminal half of Vps35 must be close to the membrane. The flexibility of the cargo-recognition complex suggests that it could adapt to the shape of curved tubulovesicular membranes and so bind multiple membrane-embedded cargo molecules. The curvature and contour length of the SNX dimer can be estimated from the related structure of the amphiphysin BAR domain[30], which has a contour length of 150 Å along its inner-membrane binding surface. The cargo recognition complex has at least three distinct points of interaction with the SNX dimer. Their contour lengths are similar, suggesting that they could form curved side-by-side arrays at a 1:1 ratio (Fig. 4b). Although speculative, this conceptual model helps lay a foundation for further analysis of the assembly of the full retromer complex on cargo-bearing vesicles.

The clathrin and COPII coats, which direct the budding of roughly spherical vesicles, have been the principal models for the structural
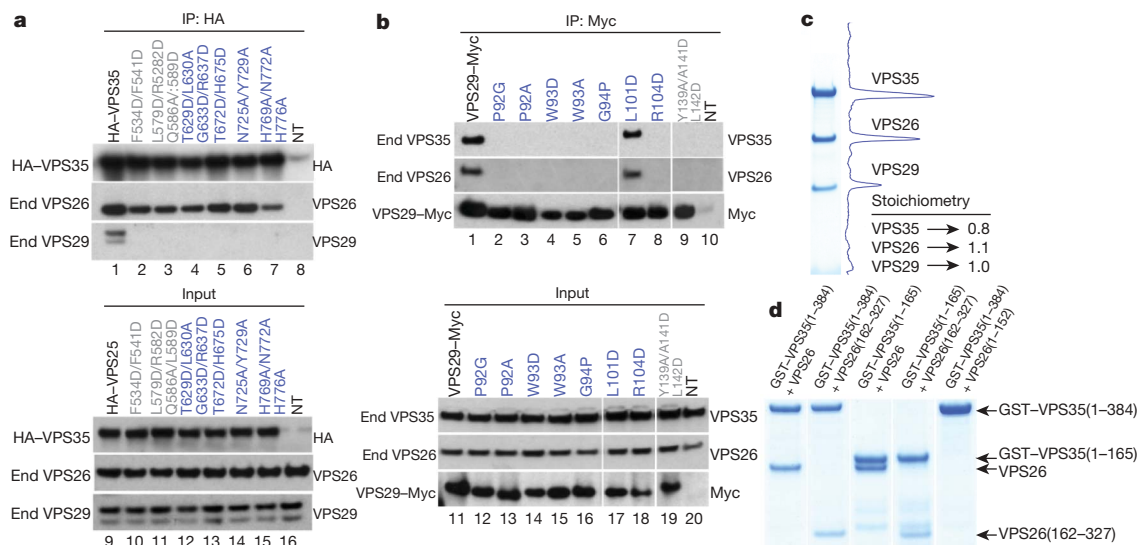

Figure 2 | Assembly of the cargo-recognition complex. a, Mutations in the VPS29-interaction surface of VPS35 block assembly *in vivo*. Lysates from HeLa cells transfected with cDNAs encoding wild-type or mutant forms of HA-tagged VPS35, or lysates from non-transfected cells (lanes 8 and 16, NT), were subjected to immunoprecipitation (IP) using mouse monoclonal antibody to the HA epitope. Lysates (2% of total, lanes 9–16) and immunoprecipitates (lanes 1–8) were analysed by SDS–PAGE and immunoblotting with antibodies to the HA tag, VPS26 and VPS29. Coloured text matches the colour scheme in Fig. 1. **b**, Mutations in the VPS35-interaction surface of VPS29 abrogate assembly, with the exception of L101D. Lysates from HeLa cells transfected with cDNAs encoding wild-type or mutant VPS29–Myc, or lysates from non-transfected cells (lanes 10 and 20, NT), were subjected to immunoprecipitation using a mouse monoclonal antibody to the Myc epitope. Lysates (5% of total, lanes 11–20) and

immunoprecipitates (lanes 1–10) were analysed by SDS–PAGE and immunoblotting with antibodies to VPS35, VPS26 and the Myc tag. **c**, The cargo-recognition complex is an equimolar complex of VPS26, VPS29 and VPS35. Gels were stained with Coomassie blue (Simply Blue SafeStain; Invitrogen). Image acquisition and analysis was done in an Epichemi Darkroom (UVP BioImaging Systems) using Labworks 4.5 software. The integrated area from each peak was normalized to the calculated molecular mass for each protein, and the value was determined relative to VPS29. **d**, Coomassie-stained gel showing that the N terminus of VPS35 forms a stable subcomplex with the C-terminal lobe of VPS26. For each lane, GST-tagged VPS35 fragments and untagged VPS26 or its fragments were co-expressed in *Escherichia coli* and purified by glutathione-Sepharose chromatography.
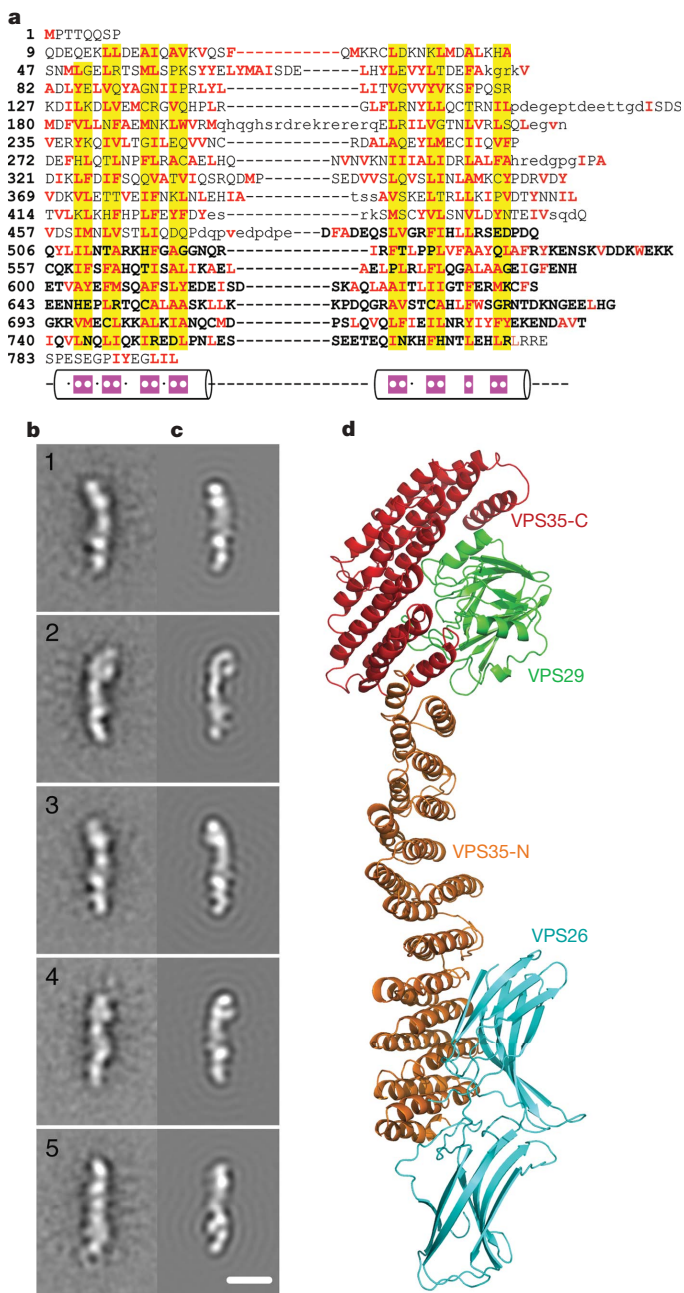
dissection of coat complexes[20–22]. These two coats consist of an inner layer (the clathrin adaptors and the Sec23–Sec24 complex, respectively) that binds to cargo proteins and lipids, and an outer layer that forms an organizing cage (composed of clathrin and the Sec13–Sec31 complex, respectively) but does not directly contact cargo or lipids. Retromer has extensive structural similarities to these classical systems: VPS26 looks like the clathrin adaptor β-arrestin[16], whereas VPS35 has an α-solenoid fold similar to the AP trunk domains and to the clathrin heavy chain itself. On the other hand, retromer contrasts sharply in that the inner layer functions of the classical coats—binding to cargo and membrane lipids—are divided between the cargo recognition and SNX complexes, and are not shared by one inner complex. Neither the cargo recognition nor the SNX complex appears to serve as an outer coat, and it is unclear if an outer coat exists for this system. The apparent lack of an outer coat in this system is consistent with the absence of a classical electron-dense coat on vesicles of retrograde trafficking[2]. The putative single-layer 'cage free'

Figure 3 | Structural analysis of the complete cargo-recognition complex. **a**, Alignment of human VPS35 repeats. Columns of residues that are likely to form the hydrophobic core of the structure have a yellow background. Apolar residues are in red. Cylinders show the α-helical regions. Magenta boxes within the α-helices indicate the positions of apolar residues in the consensus sequences. The C-terminal region of the crystal structure is in bold. Loop regions predicted on the basis of the multiple sequence alignment of VPS35 proteins are shown by lower-case letters. **b**, Averaged images of the cargo-recognition complex from negative stain electron microscopy (Supplementary Fig. 6). The images were obtained by multivariate statistical analysis with reference-free alignment. The number of images per class 1 to 5 is respectively 288, 362, 367, 206 and 327. **c**, Corresponding projections, limited to 25 Å resolution, of the cargo-recognition complex model shown in **d** (C-terminal crystallized region of VPS35, red; VPS29, green; VPS26, cyan; and N-terminal modelled region of VPS35, orange). Each image in **c** is oriented such that VPS35–VPS29 corresponds always to the top part of the image and VPS26 to the bottom part, as in **d**. The correlation coefficient between the electron microscopy class average and the corresponding model projection for class 1 to 5 is respectively 0.79, 0.70, 0.86, 0.75 and 0.73. Scale bar in **c**, 100 Å.
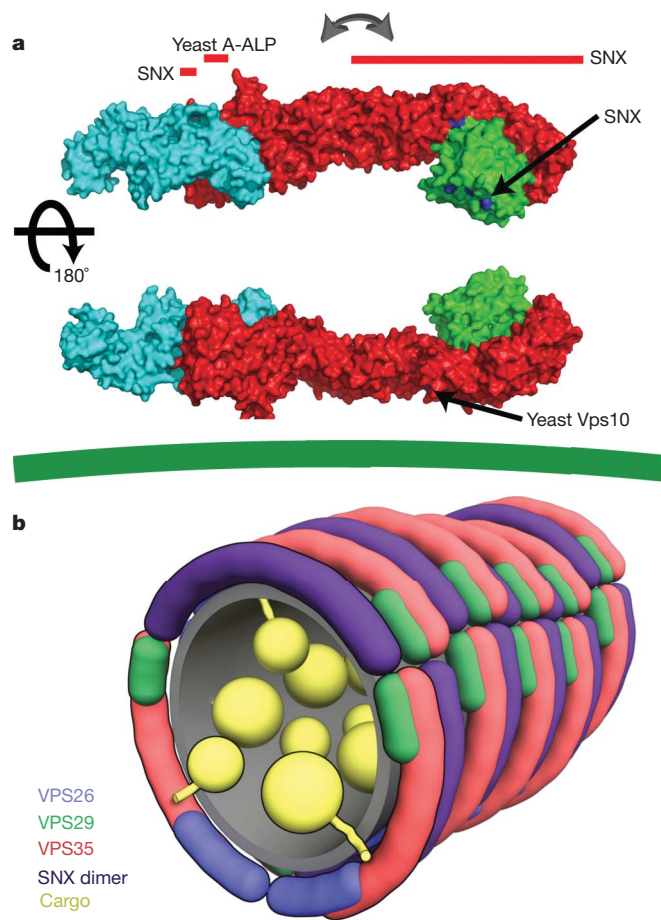
Figure 4 | Integration of cargo and targeting signals by the cargo-recognition complex. **a**, The VPS26–VPS29–VPS35 complex is predicted to align roughly parallel to the membrane (green line at bottom), such that its multiple SNX[4,15] and cargo-binding sites[25] cooperatively interact. The arrows mark the central region about which VPS35 is proposed to flex so as to interact with cargo embedded in curved membranes. Binding sites that have been mapped to individual residues within crystallized components are coloured dark blue. Binding sites that have been mapped to regions of VPS35 or to as yet non-crystallized portions of VPS35 are marked by red bars aligned with the region of interest. Binding sites for yeast cargo proteins are not necessarily conserved in human VPS35; however, the overall architecture of the yeast and other orthologous complexes is proposed to be very similar to the human complex. **b**, Schematic rendering of a speculative model for the retromer coat on a tubular vesicle, coloured as above, with the SNX dimer in purple.

organization of the system seems to be ideally designed for the coating of tubular vesicles of variable length. The rough cylindrical symmetry and ill-defined endpoints present very different geometric constraints compared with the quasi-spherical vesicles of endocytosis and the secretory pathway, and thus demand different solutions.

METHODS SUMMARY

The crystal structure of the VPS29–VPS35 C-terminal subcomplex was determined by single anomalous dispersion at the Se edge and refined to an *R*-free value of 0.278 at 2.8 Å resolution.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Seaman, M. N. Recycle your receptors with retromer. *Trends Cell Biol.* **15**, 68–75 (2005).
2. Bonifacino, J. S. & Rojas, R. Retrograde transport from endosomes to the trans-Golgi network. *Nature Rev. Mol. Cell Biol.* **7**, 568–579 (2006).
3. Seaman, M. N. J., McCaffery, J. M. & Emr, S. D. A membrane coat complex essential for endosome-to-Golgi retrograde transport in yeast. *J. Cell Biol.* **142**, 665–681 (1998).
4. Haft, C. R. *et al.* Human orthologs of yeast vacuolar protein sorting proteins Vps26, 29, and 35: Assembly into multimeric complexes. *Mol. Biol. Cell* **11**, 4105–4116 (2000).
5. Arighi, C. N., Hartnell, L. M., Aguilar, R. C., Haft, C. R. & Bonifacino, J. S. Role of the mammalian retromer in sorting of the cation-independent mannose 6-phosphate receptor. *J. Cell Biol.* **165**, 123–133 (2004).
6. Carlton, J. *et al.* Sorting nexin-1 mediates tubular endosome-to-TGN transport through coincidence sensing of high-curvature membranes and 3-phosphoinositides. *Curr. Biol.* **14**, 1791–1800 (2004).
7. Seaman, M. N. Cargo-selective endosomal sorting for retrieval to the Golgi requires retromer. *J. Cell Biol.* **165**, 111–122 (2004).
8. Verges, M. *et al.* The mammalian retromer regulates transcytosis of the polymeric immunoglobulin receptor. *Nature Cell Biol.* **6**, 763–769 (2004).
9. Coudreuse, D. Y. M., Roel, G., Betist, M. C., Destree, O. & Korswagen, H. C. Wnt gradient formation requires retromer function in Wnt-producing cells. *Science* **312**, 921–924 (2006).
10. Prasad, B. C. & Clark, S. G. Wnt signaling establishes anteroposterior neuronal polarity and requires retromer in *C. elegans. Development* **133**, 1757–1766 (2006).
11. Strochlic, T. I., Setty, T. G., Sitaram, A. & Burd, C. G. Grd19/Snx3p functions as a cargo-specific adapter for retromer-dependent endocytic recycling. *J. Cell Biol.* **177**, 115–125 (2007).
12. Small, S. A. *et al.* Model-guided microarray implicates the retromer complex in Alzheimer's disease. *Ann. Neurol.* **58**, 909–919 (2005).
13. Rojas, R., Kametaka, S., Haft, C. R. & Bonifacino, J. S. Interchangeable but essential functions of SNX1 and SNX2 in the association of retromer with endosomes and the trafficking of the mannose 6-phosphate receptors. *Mol. Cell. Biol.* **27**, 1112–1124 (2007).
14. Wang, D. *et al.* Crystal structure of human vacuolar protein sorting protein 29 reveals a phosphodiesterase/nuclease-like fold and two protein-protein interaction sites. *J. Biol. Chem.* **280**, 22962–22967 (2005).
15. Collins, B. M., Skinner, C. F., Watson, P. J., Seaman, M. N. & Owen, D. J. Vps29 has a phosphoesterase fold that acts as a protein interaction scaffold for retromer assembly. *Nature Struct. Mol. Biol.* **12**, 594–602 (2005).
16. Shi, H., Rojas, R., Bonifacino, J. S. & Hurley, J. H. The retromer subunit Vps26 has an arrestin fold and binds Vps35 through its C-terminal domain. *Nature Struct. Mol. Biol.* **13**, 540–548 (2006).
17. Damen, E., Krieger, E., Nielsen, J. E., Eygensteyn, J. & van Leeuwen, J. E. M. The human Vps29 retromer component is a metallo-phosphoesterase for a cation-independent mannose 6-phosphate receptor substrate peptide. *Biochem. J.* **398**, 399–409 (2006).
18. Collins, B. M., McCoy, A. J., Kent, H. M., Evans, P. R. & Owen, D. J. Molecular architecture and functional model of the endocytic AP2 complex. *Cell* **109**, 523–535 (2002).
19. Heldwein, E. E. *et al.* Crystal structure of the clathrin adaptor protein 1 core. *Proc. Natl Acad. Sci. USA* **101**, 14108–14113 (2004).
20. Ybe, J. A. *et al.* Clathrin self-assembly is mediated by a tandemly repeated superhelix. *Nature* **399**, 371–375 (1999).
21. Fotin, A. *et al.* Molecular model for a complete clathrin lattice from electron cryomicroscopy. *Nature* **432**, 573–579 (2004).
22. Fath, S., Mancias, J. D., Bi, X. & Goldberg, J. Structure and organization of coat proteins in the COPII cage. *Cell* **129**, 1251–1252 (2007).
23. Misra, S., Puertollano, R., Kato, Y., Bonifacino, J. S. & Hurley, J. H. Structural basis for acidic-cluster-dileucine sorting-signal recognition by VHS domains. *Nature* **415**, 933–937 (2002).
24. Shiba, T. *et al.* Structural basis for recognition of acidic-cluster dileucine sequence by GGA1. *Nature* **415**, 937–941 (2002).
25. Nothwehr, S. F., Bruinsma, P. & Strawn, L. A. Distinct domains within Vps35p mediate the retrieval of two different cargo proteins from the yeast prevacuolar/endosomal compartment. *Mol. Biol. Cell* **10**, 875–890 (1999).
26. Yang, J. *et al.* Molecular basis for TPR domain-mediated regulation of protein phosphatase 5. *EMBO J.* **24**, 1–10 (2005).
27. Groves, M. R. & Barford, D. Topological characteristics of helical repeat proteins. *Curr. Opin. Struct. Biol.* **9**, 383–389 (1999).
28. Conti, E. & Izaurralde, E. Nucleocytoplasmic transport enters the atomic age. *Curr. Opin. Cell Biol.* **13**, 310–319 (2001).
29. Nothwehr, S. F., Ha, S. A. & Bruinsma, P. Sorting of yeast membrane proteins into an endosome-to-Golgi pathway involves direct interaction of their cytosolic domains with Vps35p. *J. Cell Biol.* **151**, 297–310 (2000).
30. Peter, B. J. *et al.* BAR domains as sensors of membrane curvature: The amphiphysin BAR structure. *Science* **303**, 495–499 (2004).

METHODS

Protein expression and purification. DNA encoding full-length human VPS29 and residues 476–780 of human VPS35 (C terminus) was amplified using the polymerase chain reaction (PCR). The products were inserted into pmr101A (modified pmr101; American Type Culture Collection) and pGEX-KG, respectively. VPS35 was fused to an N-terminal GST tag followed by a thrombin cleavage site, and VPS29 was untagged. The plasmids were co-transformed into *Escherichia coli* BL21 (DE3) star (Invitrogen) and protein was expressed at 15 °C for 12 h by induction with 1 mM isopropylthiogalactoside (IPTG) when cells reached an A_{600} of 0.8. Cells were lysed by high-pressure homogenization in 50 mM Tris (pH 8.0), 1 M NaCl and 10 mM dithiothreitol (DTT). The complex was isolated using glutathione-Sepharose TM 4B (Amersham Biosciences) affinity chromatography. The fusion protein was cleaved with thrombin while bound to the matrix in 50 mM Tris (pH 8.0), 50 mM NaCl and 10 mM DTT. The subcomplex was further purified by ion exchange (5 ml HiTrap Q FF) and gel filtration (Superdex 200 16/60) (GE Healthcare) chromatography. Fractions were pooled, concentrated to 2 mg ml^{-1} in 50 mM Tris (pH 8.0), 1 M NaCl, 10 mM DTT, and stored at −80 °C. Full-length human VPS35, VPS26A and VPS29 cDNAs were amplified by PCR, and the products were subcloned into pGST-parallel2 vector (VPS35) and pmr101A vectors (VPS26A and VPS29), respectively. The VPS29 and VPS35 expression plasmids were co-transformed into BL21 (DE3) star whereas VPS26A was transformed separately. Expression was induced as above except with 0.3 mM IPTG. Cells from both lines were harvested together and lysed by high-pressure homogenization in 50 mM Tris (pH 8.0), 150 mM NaCl and 10 mM DTT. The full complex was then purified as described above. Human VPS29 was expressed and purified as above.

VPS26–VPS35 fragment interactions. For mapping VPS35–VPS26 interaction, the appropriate regions of human VPS35 were cloned into parallel GST2 vectors[31], using a standard PCR-based cloning strategy, from cDNA encoding VPS35. Subsequently, VPS26A and GST–VPS35 constructs were amplified by PCR to generate cassettes containing the Shine-Delgarno translational start signal[32] and cloned directly into the polycistronic pST39 vector[33]. Proteins were expressed as for the intact complex and purified by glutathione-Sepharose affinity chromatography as described above.

Crystallization and data collection. Crystals of the retromer VPS29–VPS35 subcomplex were grown at 18 °C using hanging-drop vapour diffusion by mixing 2 μl of 16 mg ml^{-1} protein solution (50 mM Tris pH 8.0, 1 M NaCl, 15 μM glycyl-glycyl-glycine) with 4 μl of 150 mM potassium sodium tartrate and 20% PEG 3350. Crystals were cryo-protected with 15% glycerol and flash frozen in liquid nitrogen. SAD experimental data at $\lambda = 0.9793$ Å were collected at 95 K at APS SER-CAT Beamline 22-ID at the Se edge, and processed using HKL2000 (HKL Research).

Structure determination and refinement. Twelve of fourteen expected selenium sites were located and refined using SOLVE[34], and initial phases were calculated at 3.1 Å. An interpretable electron density map was obtained after density modification and phase extension to 2.8 Å with RESOLVE. This map was used to place VPS29 in the subcomplex structure and the VPS35 C terminus was built by hand in COOT[35]. Subsequent rounds of model adjustment, simulated annealing and thermal factor refinement were performed with CNS[36]. During the final refinement stages, water molecules were inserted into the protein model. Statistics are summarized in Supplementary Table 1. The final atomic model is a dimer of the retromer subcomplex in the asymmetric unit, where chains A and B correspond to VPS29 and chains C and D correspond to VPS35 C terminus. This model includes two glycerol molecules and 53 water molecules. 88.4 % of the residues are in the most allowed region of the Ramachandran plot. Residues 41 and 115 of VPS29 are the only residues in the disallowed region of the Ramachandran plot, and are in the same conformations as in the structure of VPS29 previously determined in isolation. No electron density was observed for residues 679–694 or 734–738 in chain D. Figure 1d was coloured using the HotPatch server[37].

In vitro phosphatase assays. Purified proteins were dialysed against 150 mM NaCl, 50 mM Tris/HCl pH 7.0 buffer. Each protein or protein complex was assayed at a final concentration of 17.5 μM in the presence of 1,500 μM of serine-phosphorylated peptide SFHDDpSDEDLLHI (>98% purity; New England Peptide LLC) and 50 μM ZnCl$_2$. The volume for each reaction was adjusted to 100 μl with serine/threonine phosphatase assay buffer (50 mM Tris/HCl, pH 7.0, and 100 μM CaCl$_2$) (Upstate Biotechnology). EDTA was added to duplicate reactions at a final concentration of 5 mM. As a positive control, 0.25 μM of shrimp alkaline phosphatase (Roche) was used under the same conditions. The reactions were incubated at 37 °C for 1 h or 16 h. From each sample, 30 μl was added to 60 μl of Biomol green (Biomol) and the amount of free phosphate was determined from the absorbance at 630 nm.

Bioinformatics of metallophosphoesterases. The alignment of 5,307 putative metallophosphoesterases was downloaded from the Pfam database[38] and residue conservation at the designated positions was analysed with a locally written script.

Recombinant DNA constructs for cell biology studies. An *Eco*RI–*Bam*HI fragment (including a newly introduced start codon in position 1, and the replacement of a stop codon in position 183 for a tyrosine residue) encoding human VPS29 was obtained by PCR amplification from the VPS29-pGADT7 construct[16]. This fragment was later cloned into the corresponding sites of the pcDNA3.1+/Myc-His vector (Invitrogen). An *Mfe*I–*Not*I fragment encoding VPS35 (2–796) was amplified by PCR from the VPS35-pGBKT7 construct and cloned into the respective sites of the pCI-HA$_3$ vector[39]. This construct was used as a template for the PCR amplification of the (HA)$_3$–VPS35 fragment that was subsequently cloned into the pEF6/V5-His TOPO TA vector (Invitrogen). Mutagenesis of both the VPS29–Myc-pcDNA3.1 and (HA)$_3$–VPS35-pEF6/V5 constructs was performed using the QuickChange site-directed mutagenesis kit (Stratagene).

Immunoprecipitation. HeLa cells (American Type Culture Collection) were cultured on 35-mm dishes as described previously[13]. Once the cells reached 80–90% confluency, they were transfected with 4 μg of wild-type or mutant (HA)$_3$–VPS35-pEF6/V5 plasmids using Lipofectamine 2000 (Invitrogen). After 36 h, cells lysates were prepared and pre-cleared as previously described[16]. Pre-cleared lysates were then incubated for 2 h at 4 °C with 30 μl of protein-G-Sepharose beads bound to mouse monoclonal antibody to HA (clone HA. 11; Covance). Immunoprecipitates were washed and then subjected to SDS–PAGE and immunoblot analysis using antibodies to VPS29, VPS26 and HA, as described previously[13,16]. Immunoprecipitation of HeLa cells expressing either wild-type or mutant VPS29–Myc-pcDNA3.1 was performed as described above but with the difference that cell lysates were incubated with protein-G-Sepharose beads conjugated to mouse monoclonal antibody to Myc (9E10; Covance). Antibodies to VPS35 and Myc used for immunoblotting have been described previously[13].

Immunofluorescence labelling and microscopy. HeLa cells grown on coverslips were transfected with wild-type and mutant VPS29–Myc-pcDNA3.1 or wild-type and mutant (HA)$_3$–VPS35-pEF6/V5 constructs with Lipofectamine 2000 as indicated above. At 36 h after transfection, cells were washed, fixed and labelled with mouse monoclonal antibody to Myc (9E10), rat monoclonal antibody to HA (3F10) and rabbit polyclonal antibody to VPS26, and the respective fluorescently labelled secondary antibodies. Imaging was performed using an epifluorescence microscope (AX10; Carl Zeiss Microimaging, Inc) with a Plan-Apochromatic ×63 oil objective and the appropriate filter combination.

Electron microscopy and image processing. VPS35–VPS29–VPS26 retromer complexes (10 μg ml^{-1}) were negatively stained with 1% uranyl formate and observed on a Philips CM120 microscope with LaB6 illumination. Micrographs, recorded at 60,000 magnification on Kodak SO-163 film, were digitized on a Nikon 9000 scanner and binned twofold, giving a pixel size of 2.1 Å. Particle picking, CTF correction by phase flipping, and other pre-processing operations, were carried out with Bsoft[40]. 4,331 particles, low-pass-filtered at 15 Å, were sorted by reference-free alignment as implemented in EMAN[41], then classified in SPIDER[42] with multiple rounds of correspondence analysis, hierarchical clustering by the Ward criteria, and multi-reference alignment until stable classes were obtained. In the last iteration, thirteen classes, representing the major trends within the data set (3,815 particles), were individually averaged. Their resolutions, determined by Fourier ring correlation at 0.5 cutoff, were between 24 and 26 Å.

Identification of VPS35 repeats. We probed the N-terminal part of VPS35 (residues 1–480) for additional copies of repeats by applying a sensitive sequence profile method and pftools package[43]. First, we constructed the profile from structure-based alignments of the repeats from the C-terminal part of VPS35. The profile spanned three VPS35 repeats to increase the selectivity of the search. This is in line with the observations that a single repeat would be unlikely to form a stable structure on its own and that α-helical solenoid proteins with tandem arrays of repeats. The best match produced by this profile (positions 459–592) had an *E*-value <0.0001. The *E*-values of the matches were calculated as described[44]. Subsequent examination of this match found mostly hydrophobic residues in positions that should be apolar because the corresponding residues of known VPS35 structure have interior locations. The newly identified repeats were included in the initial structure-based alignment and the second profile was constructed. The profile identified three additional repeats in the N-terminal region (positions 272–413) with an *E*-value <0.000001. For the remainder of VPS35, the repeat profiles produced a few weaker matches with *E*-values between 0.01 and 1. At the same time, several homologous VPS35 sequences were aligned and this allowed us to determine whether the residues conserved in the aligned repeats are also conserved between homologues. The alignment of VPS35 homologues also identified sites of insertions that probably correspond to loops in the

three-dimensional structures. This information was used to generate the alignment of the remaining repeats with the weak matches.

Structural interpretation of electron microscopy. A model of residues 1–482 of VPS35 as a right-handed α-solenoid of 21 helices was constructed using the crystal structure of residues 483–780 as a template. The model was built manually, taking two copies of the 483–780 fragment containing a total of 26 helices as the starting point. The two fragments were manually aligned end to end with each other and with the VPS29–VPS35 subcomplex in a linear fashion as suggested by the processed electron microscopy images. Helical pairs at the N and C termini of each fragment were overlapped in order to maintain the pitch of the solenoid. The overlapping helices were then removed from the model, leaving a total of 21 modelled helices. Because the electron microscopy images offer no guidance for side-chain placement, no effort was made to assign the correct amino acid residues to the model. VPS26 was docked onto the N-terminal 165 residues of VPS35 such that its C-terminal lobe and in particular the loop from residues 235–246 (ref. 16) were in contact with each other. The curvature of the model was manually adjusted such that the overall cargo recognition complex model had a 210 Å end-to-end distance. The mean contour length of 0.26 Å per residue observed in the crystallized portion of VPS35 was maintained in the model.

For comparison with the electron microscopy class averages, the molecular model was band-limited to 25 Å resolution with Bsoft and two-dimensional projections were calculated, corresponding to a set of equally spaced side views. Finally, a CTF corresponding to the defocus condition of the micrographs was applied. The coordinates of the molecular model were then modified manually as follows to improve the agreement with the electron microscopy data. No modifications were made to the crystallographically determined structures of VPS35-C–VPS29 and VPS26. The relative position of the VPS35–VPS29 subcomplex versus VPS26 interacting with the N-terminal 165 residues of VPS35 was progressively adjusted by small increments and the modelled portion of VPS35 shifted to maintain continuity. Then, a new set of two-dimensional model projections was generated and compared with the electron microscopy data, both visually and computationally (by calculating correlation coefficients). Finally, the placement of VPS26 on the N-terminal region of VPS35 was also modified manually in a way that kept the total length of the model at 210 Å (the value consistently observed in the electron microscopy analysis) and maintained contact between VPS35-N and the VPS26 loop from residues 235–246, and evaluated in the same way. After these adjustments, disconnected loops were manually

reconnected in O[45], and the structure was energy-minimized with CNS[36] to remove gaps and steric clashes. The consistency between the final model and the electron microscopy class average was estimated by computing the average Fourier ring cross-correlation curve as a function of resolution, giving a 29 Å limit for a 0.3 cutoff (Supplementary Fig. 8).

31. Sheffield, P., Garrard, S. & Derewenda, Z. Overcoming expression and purification problems of RhoGDI using a family of "parallel" expression vectors. *Protein Expr. Purif.* **15**, 34–39 (1999).
32. Hierro, A., Kim, J. & Hurley, J. H. Polycistronic expression and purification of the ESCRT-II endosomal trafficking complex. *Methods Enzymol.* **403**, 322–332 (2005).
33. Tan, S. A modular polycistronic expression system for overexpressing protein complexes in *Escherichia coli. Protein Expr. Purif.* **21**, 224–234 (2001).
34. Terwilliger, T. C. & Berendzen, J. Automated MAD and MIR structure solution. *Acta Crystallogr. D* **55**, 849–861 (1999).
35. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
36. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
37. Pettit, F. K., Bare, E., Tsai, A. & Bowie, J. U. HotPatch: a statistical approach to finding biologically relevant features on protein surfaces. *J. Mol. Biol.* **369**, 863–879 (2007).
38. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **30**, 276–280 (2002).
39. Martina, J. A., Moriyama, K. & Bonifacino, J. S. BLOC-3, a protein complex containing the Hermansky-Pudlak syndrome gene products HPS1 and HPS4. *J. Biol. Chem.* **278**, 29376–29384 (2003).
40. Heymann, J. B. & Belnap, D. M. Bsoft: image processing and molecular modeling for electron microscopy. *J. Struct. Biol.* **157**, 3–18 (2007).
41. Ludtke, S. J., Baldwin, P. R. & Chiu, W. EMAN: semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.* **128**, 82–97 (1999).
42. Frank, J. *et al.* SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *J. Struct. Biol.* **116**, 190–199 (1996).
43. Bucher, P., Karplus, K., Moeri, N. & Hofmann, K. A flexible motif search technique based on generalized profiles. *Comput. Chem.* **20**, 3–23 (1996).
44. Hofmann, K. & Bucher, P. The FHA domain: a putative nuclear signalling domain found in protein kinases and transcription factors. *Trends Biochem. Sci.* **20**, 347–349 (1995).
45. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron-density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).

LETTERS

Regulation of cell cycle progression and gene expression by H2A deubiquitination

Heui-Yun Joo[1]*, Ling Zhai[1]*, Chunying Yang[1], Shuyi Nie[2], Hediye Erdjument-Bromage[3], Paul Tempst[3], Chenbei Chang[2] & Hengbin Wang[1]

Post-translational histone modifications have important regulatory roles in chromatin structure and function[1–3]. One example of such modifications is histone ubiquitination, which occurs predominately on histone H2A and H2B. Although the recent identification of the ubiquitin ligase for histone H2A has revealed important roles for H2A ubiquitination in *Hox* gene silencing[4–6] as well as in X-chromosome inactivation[7,8], the enzyme(s) involved in H2A deubiquitination and the function of H2A deubiquitination are not known. Here we report the identification and functional characterization of the major deubiquitinase for histone H2A, Ubp-M (also called USP16). Ubp-M prefers nucleosomal substrates *in vitro*, and specifically deubiquitinates histone H2A but not H2B *in vitro* and *in vivo*. Notably, knockdown of Ubp-M in HeLa cells results in slow cell growth rates owing to defects in the mitotic phase of the cell cycle. Further studies reveal that H2A deubiquitination by Ubp-M is a prerequisite for subsequent phosphorylation of Ser 10 of H3 and chromosome segregation when cells enter mitosis. Furthermore, we demonstrate that Ubp-M regulates *Hox* gene expression through H2A deubiquitination and that blocking the function of Ubp-M results in defective posterior development in *Xenopus laevis*. This study identifies the major deubiquitinase for histone H2A and demonstrates that H2A deubiquitination is critically involved in cell cycle progression and gene expression.

To identify the deubiquitinase for histone H2A, we developed an *in vitro* deubiquitination assay using ubiquitinated H2A (ubH2A)-containing mononucleosomes as substrates (Fig. 1a and Supplementary Fig. 1a, b)[6,9,10]. When these nucleosomes were incubated with HeLa nuclear protein fractions, a decrease of ubH2A levels and an increase of the released intact ubiquitin were detected, suggesting that an H2A-specific deubiquitinase is present in the corresponding fractions (Fig. 1a, lanes 2–4)[11,12]. We first focused on the nuclear extract P11 0.5 M fraction, which contains the strongest activity. Through a six-column purification scheme, we identified a polypeptide that seems to be responsible for the H2A deubiquitination activity (Fig. 1b). Silver staining (top panels) and enzymatic activity assays (middle panels) of fractions derived from the hydroxyapatite and Superose 6 columns allowed us to correlate the deubiquitinase activity with one protein band that migrates around 120 kDa in the SDS–polyacrylamide gel electrophoresis (PAGE) assay (marked with an asterisk in Fig. 1c and Supplementary Fig. 2). The relative mass of the putative H2A deubiquitinase was estimated to be approximately 500 kDa, suggesting that it functions as a homo-tetramer (Fig. 1c). Mass spectrometry analysis identified the protein as Ubp-M, a previously described ubiquitin-processing protease (data not shown)[13–15]. Western blot assays confirmed the identity of the 120-kDa protein and

revealed that the elution profiles of Ubp-M correlated with H2A deubiquitination activity (bottom panels in Fig. 1c and Supplementary Fig. 2). To verify that Ubp-M is indeed responsible for the H2A deubiquitination activity, we performed an immunodepletion assay. As shown in Fig. 1d, when Ubp-M was specifically immunodepleted from an aliquot of the hydroxyapatite column fractions, the H2A deubiquitination activity was also depleted (compare lanes 1–3 and 4–6). Furthermore, recombinant Ubp-M, which was expressed and purified from Sf9 cells, also elutes at approximately 500 kDa and is as active as native Ubp-M in H2A deubiquitination (Supplementary Fig. 3a, b). On the basis of these studies, we conclude that Ubp-M is a putative H2A deubiquitinase.

To gain insight into the function of Ubp-M, we determined its substrate preference and specificity. As shown in Fig. 2a, Ubp-M could deubiquitinate nucleosomal ubH2A, with higher efficiency for oligonucleosomes, but could not deubiquitinate histone form ubH2A (compare lanes 1–4, 5–8 and 9–12)[9]. This result indicates that Ubp-M is a nucleosome-specific deubiquitinase. To determine the substrate specificity of Ubp-M, we reconstituted mononucleosomes with Flag-tagged human H2A and Flag-tagged yeast H2B, both of which contain about 10% ubiquitinated species, together with other recombinant histones (Fig. 2b, top panels). As shown in Fig. 2b, Ubp-M deubiquitinates ubH2A and results in an increase of Flag–H2A levels in a time-dependent manner (bottom panels, lanes 1–4). In contrast, Ubp-M fails to deubiquitinate ubH2B (ubiquitinated H2B) within the same time period (Fig. 2b, bottom panel, lanes 5–8). The failure to deubiquitinate ubH2B is not due to the origin of the substrate (yeast H2B), as Ubp-M also fails to deubiquitinate nucleosomal human H2B (Supplementary Fig. 4) and the nuclear pellet P11 0.5 M fraction could deubiquitinate yeast ubH2B efficiently (Fig. 2b, bottom panel, lane 9). Collectively, such results indicate that Ubp-M is an H2A-specific deubiquitinase.

Previous studies indicated that overexpression of Ubp-M reduces the levels of cellular ubH2A; however, the physiological relevance of these observations was not clear[16]. To determine whether Ubp-M deubiquitinates H2A *in vivo*, we transfected control and *Ubp-M* short interfering (si)RNA into stable cell lines that expressed Flag–H2A and Flag–H2B, respectively (Fig. 2c). Because the levels of ubH2B are low, we co-transfected haemagglutinin-tagged ubiquitin to observe ubH2B after immunoprecipitation. Transfection of *Ubp-M* siRNA reduces the levels of Ubp-M significantly, and results in a clear increase in the levels of ubiquitinated histones (Fig. 2c, top three panels, lanes 3, 6). To determine whether the increase of ubiquitinated histones was due to ubH2A and/or ubH2B, we performed denaturing immunoprecipitation assays[17,18]. As indicated in Fig. 2c, Ubp-M knockdown results in a specific increase in the levels of

ubH2A but not ubH2B (Fig. 2c, fourth panel, and Supplementary Fig. 5). On the basis of these experiments, we conclude that Ubp-M specifically deubiquitinates histone H2A *in vivo*.

To determine the function(s) of H2A deubiquitination *in vivo*, we generated HeLa cell lines with stable knockdown of Ubp-M[4]. Knockdown of Ubp-M (>90%, protein level) results in a $50 \pm 20\%$ (mean ± s.d. of five cell lines) increase of ubH2A levels (Fig. 3a, left panels). Notably, Ubp-M knockdown cells exhibit slow proliferation rates compared with control cells (threefold slower, Fig. 3a, right panel). Further analyses indicate that Ubp-M knockdown results in a decrease in the proportion of cells at G2/M but has no effect on apoptosis (Supplementary Figs 6 and 7). To substantiate this M-phase effect, control and Ubp-M knockdown cells were synchronized at G1/S by double-thymidine treatment. After release into fresh media, cells were fixed at different time points and stained with the anti-phospho-histone H3 antibody to monitor mitotic cell progression, and with propidium iodide to monitor DNA contents. Fluorescence-activated cell sorting (FACS) analyses revealed that the M-phase cell population was reduced almost twofold in Ubp-M knockdown cells (Fig. 3b). To determine whether this effect is linked to H2A deubiquitination, we analysed the levels of ubH2A throughout the cell cycle (Fig. 3c). In control cells, ubH2A levels begin to decrease when cells enter M phase (7 h), reach very low levels when cells are in M phase (9 h), start to recover as cells exit from M phase (10–11 h), and are restored to normal levels when cells enter G1/S (12–18 h; Fig. 3c, first panel, and Supplementary Fig. 8a). In

contrast, in Ubp-M knockdown cells ubH2A levels decrease modestly as cells enter M phase, and progression through M phase is delayed (Fig. 3c, fourth panel, and Supplementary Fig. 8b). Intriguingly, fluctuations in ubH2A levels during the cell cycle are inversely correlated with the fluctuations in phospho-histone H3 levels (Fig. 3c, second and fifth panels, and Supplementary Fig. 8a, b). Immunofluorescent staining confirmed the mutually exclusive distribution of ubH2A and H3 phosphorylated at Ser 10 (H3S10ph; Supplementary Fig. 9). These results suggest that histone ubiquitination may interfere with histone H3 phosphorylation.

To determine the interplay between these two modifications, we reconstituted mononucleosomes containing ubH2A and ubH2B (Supplementary Fig. 10), and used them in Aurora-B-mediated kinase reactions. As indicated in Fig. 3d, histone ubiquitination reduced H3S10 phosphorylation (middle panel, compare lanes 4, 9 with 2, 7). Importantly, Ubp-M-mediated H2A deubiquitination restores H3S10 phosphorylation (compare lane 5 with 4), suggesting that histone deubiquitination may be a prerequisite for H3S10 phosphorylation. To determine whether histone ubiquitination affects Aurora B association with nucleosomes, we performed nucleosome pull-down assays[17]. As shown in Fig. 3e, histone ubiquitination impairs the association of Aurora B kinase with nucleosomes (middle panel, compare lanes 2, 4 with 1, 3). Because the level of endogenous ubH2B is much lower than that of ubH2A (Supplementary Fig. 5), these results strongly suggest that Ubp-M-mediated H2A deubiquitination constitutes the primary mechanism that regulates the
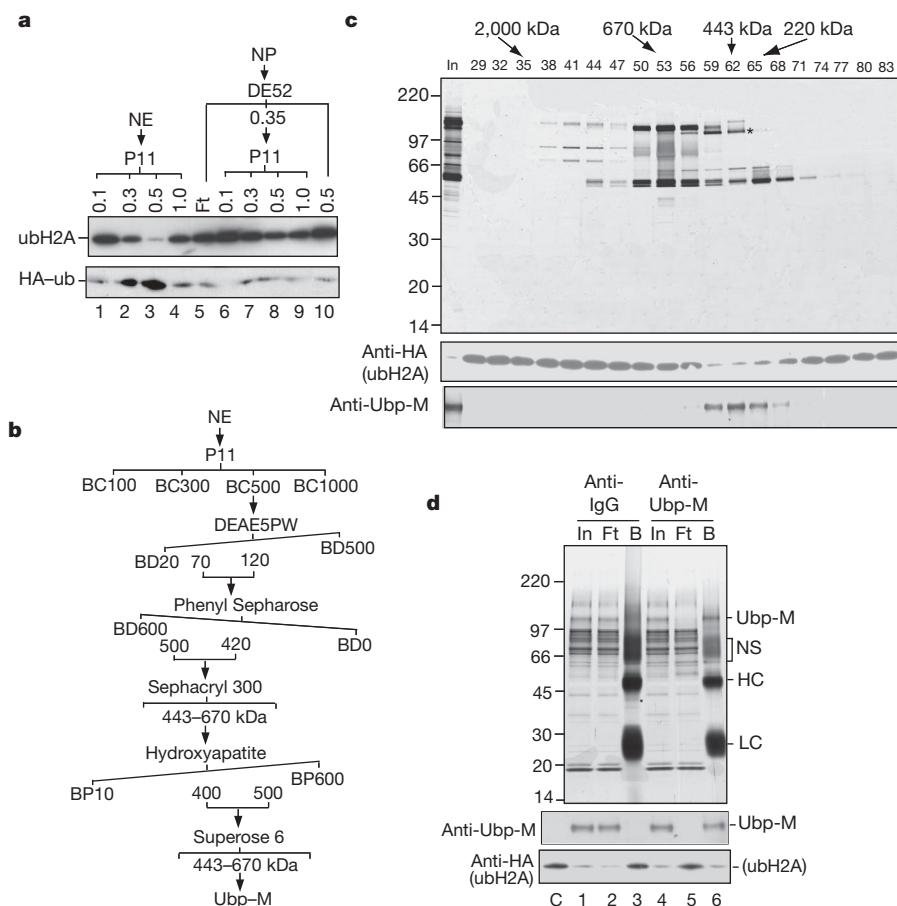

Figure 1 | Purification of the H2A-specific deubiquitinase. a, H2A deubiquitination assay with HeLa nuclear extracts (NE) and nuclear pellet (NP) proteins fractionated on DE52 and P11 columns. **b**, Schematic representation of the steps used for the purification. Numbers represent the salt concentrations (mM) at which the deubiquitinase activity elutes from the columns. **c**, Silver staining (top panel), H2A deubiquitination assay (middle panel) and western blot analysis (bottom panel) of the Superose 6 column fractions. The candidate band is indicated by an asterisk. **d**, Silver staining (top panel), western blot analysis (middle panel) and H2A deubiquitination assay (bottom panel) of input (In), flow through (Ft) and bound (B) with purified IgG and Ubp-M antibody. HC, heavy chain; LC, light chain; NS, nonspecific binding. The protein size markers (kDa) are shown on the left side of **c** and **d**.

1069

association of Aurora B kinase with nucleosomes and the subsequent H3S10 phosphorylation.

Previous studies indicated that H2A ubiquitination regulates *Hox* gene silencing[4–6]. To determine the roles of H2A deubiquitination in this process, we measured the expression of selected *Hox* genes with semi-quantitative polymerase chain reaction with reverse transcription (RT–PCR) in control and Ubp-M knockdown cells. As indicated in Fig. 4a and Supplementary Fig. 11, knockdown of Ubp-M results in a specific decrease of *HOXD10* gene expression, which could be rescued by transfection of wild type but not enzymatically inactive

(C205S) Ubp-M (Supplementary Fig. 12)[15]. To determine the role of H2A deubiquitination in *Hox* gene expression, we measured the binding of Ubp-M and Bmi1 across the *HOXD10* gene (see diagram in Fig. 4b)[6]. In control cells, Ubp-M and Bmi1 bound to two regions of *HOXD10*, one localized in the promoter region (overlapping with transcription start sites) and the other localized in 5′ regulatory regions (1.7 kb upstream of the transcription start site; Fig. 4b, top two panels, lanes 1–8). In Ubp-M KD1 knockdown cells, the binding of Ubp-M to both regions of *HOXD10* was markedly decreased whereas the binding of Bmi1 was not changed (Fig. 4b, top two panels, lanes 9–16). Chromatin double immunoprecipitations (anti-Flag immunoprecipitation followed by anti-HA immunoprecipitation) revealed that the levels of ubH2A at Ubp-M binding regions increased markedly after Ubp-M knockdown (Fig. 4b, fourth and fifth panels). Taken together, we conclude that Ubp-M-mediated H2A deubiquitination regulates gene expression.

Because Ubp-M regulates *Hox* gene expression, we investigated whether Ubp-M regulates body patterning in *Xenopus laevis*. The *Xenopus* homologue of Ubp-M is recognized by our antibody and has H2A deubiquitination activity (Supplementary Fig. 13a, b).
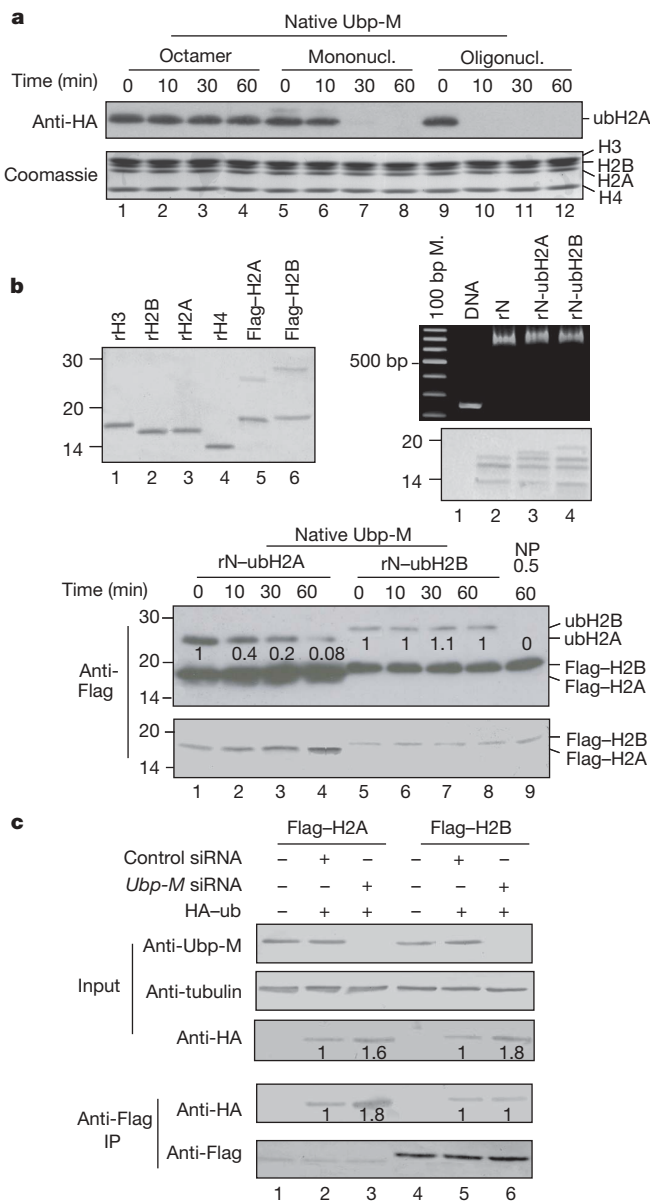

Figure 2 | Substrate preference and specificity of Ubp-M. a, Ubp-M-mediated deubiquitination reaction with substrates as indicated at the top. The amount of substrates is shown in the bottom panel. **b,** Ubp-M-mediated deubiquitination reaction with nucleosomes reconstituted with Flag–H2A and Flag–H2B as substrates (bottom panels). Top panels show nucleosome reconstitution with Flag–H2A and Flag–H2B. The protein size markers (kDa) are shown on the left side of the panels. **c,** Ubp-M is specific for H2A deubiquitination *in vivo*. HeLa cells stably expressing Flag–H2A (lanes 1–3) and Flag–H2B (lanes 4–6) were transfected with the indicated plasmid and siRNA. Western blot analysis was performed with total cellular proteins (top three panels) and anti-Flag immunoprecipitate under denaturing conditions (bottom two panels). Numbers in **b** and **c** represent the relative protein levels as quantified by Image J Software.
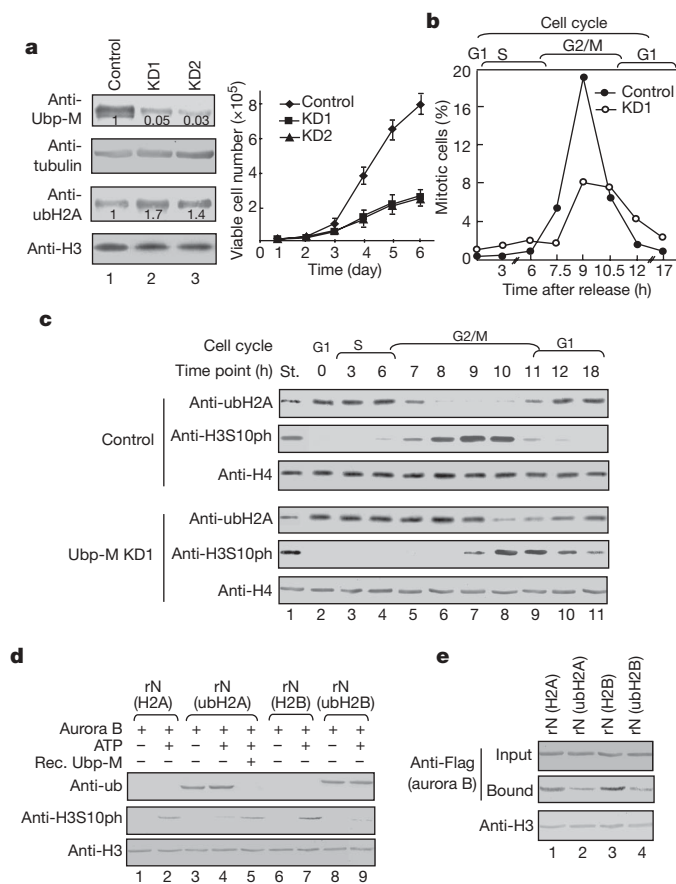

Figure 3 | Ubp-M-mediated H2A deubiquitination is required for cell cycle M phase progression. a, Ubp-M knockdown of HeLa cells results in an increase of cellular ubH2A levels (left) and slow cell growth rate (right panel). Numbers represent the relative protein levels as quantified by Image J Software. Data are represented as mean ± s.d. of three experiments. **b,** Ubp-M knockdown decreases the M-phase cell population. **c,** Western blot analyses of ubH2A and H3S10ph levels in synchronized control (top three panels) and Ubp-M KD1 (bottom three panels) cells throughout the cell cycle. Experiments were repeated three times and a representative result is shown. **d,** H2A ubiquitination inhibits H3S10 phosphorylation. Aurora B kinase reactions (middle panel) with reconstituted nucleosomes as indicated along the top of the panel. **e,** Histone ubiquitination interferes with the association between Aurora B and nucleosomes. Nucleosomes, as indicated, were incubated with Flag-tagged Aurora B kinase and the nucleosome-associated Aurora B was determined by western blot assay.

When equal amounts of purified IgG and Ubp-M antibody were injected into two-cell-stage *Xenopus* embryos, Ubp-M antibody, but not IgG, induced a homeotic transformation (enlarged cement gland and reduced posterior structures; Fig. 4c, middle panels, and Supplementary Fig. 14). The phenotype is specific, as pre-incubation of Ubp-M antibody with the Ubp-M antigen abolishes these morphological changes (Fig. 4c, bottom panels). *In situ* hybridization reveals that *Hoxd10* expression in the posterior region of the embryos is greatly reduced when Ubp-M activity is blocked by injection of the antibody (Fig. 4d). Taken together, we conclude that Ubp-M

Figure 4 | Ubp-M-mediated H2A deubiquitination regulates *Hox* gene expression. a, Semi-quantitative RT–PCR analysis of *HOXD10* expression in HeLa cells as indicated above. *GAPDH* was used as a control. **b**, ChIP assays of *HOXD10* gene binding in control (lanes 1–8) and Ubp-M knockdown (lanes 9–16) cells. A diagram of the *HOXD10* gene is illustrated at the top of the panels. The two exons and the regions that were analysed are indicated. **c**, Control uninjected *Xenopus* embryos and embryos injected with the indicated reagents are shown (side view) with the head to the left. **d**, *In situ* hybridization of *Hoxd10* expression in uninjected embryos (top left panel) or in embryos injected with IgG (top right panel) or anti-Ubp-M antibody (bottom panels).

regulates embryonic anterior–posterior patterning through the regulation of *Hox* gene expression.

Our data reveal two independent functions of Ubp-M-mediated H2A deubiquitination: (1) to regulate gene expression in G1/S phase, and (2) to facilitate cell cycle M-phase progression. Although we do not know how Ubp-M is recruited to specific regions and how Ubp-M is activated during M phase, we found that Ubp-M is phosphorylated and that the phosphorylation status of Ubp-M correlates with *in vivo* H2A deubiquitination and cell cycle progression (Supplementary Fig. 15)[15]. Further investigations will determine whether phosphorylation contributes to the regulation of Ubp-M. In addition, H2A ubiquitination has been implicated in X-chromosome inactivation[7,8], and further experiments will determine whether Ubp-M is involved in this process. Nonetheless, the identification of the major H2A deubiquitinase should greatly improve our understanding of histone ubiquitination in chromatin regulation.

METHODS SUMMARY

Ubp-M was purified from HeLa cell nuclear extracts by conventional chromatography and identified by matrix-assisted laser desorption/ionization reflectron time-of-flight (MALDI-reTOF) mass spectrometry. Recombinant Ubp-M was purified by a combination of immunoaffinity and gel filtration. *In vitro* histone deubiquitination activity was determined by western blot analysis. Native ubH2A-containing mononucleosomes were purified by tandem immunoprecipitation[6]. Mono- and oligonucleosomes were prepared by a sucrose gradient and histone octamers were purified from these nucleosomes with a hydroxyapatite column[9]. Mononucleosomes were reconstituted by serial salt dilutions[17]. Flag-tagged human H2A and yeast H2B were purified by anti-Flag immunoprecipitation[17,19]. ubH2A and ubH2B were purified from gel strips containing ubH2A and ubH2B by electro-elution. Ubp-M knockdown was achieved by RNA interference[4,17]. Effects of Ubp-M knockdown on histone ubiquitination and H3S10ph were determined by western blot analysis. Effects of Ubp-M knockdown on cell cycle progression were determined by FACS analysis. The distribution of ubH2A and H3S10ph was revealed by immunofluorescence staining. Cell synchronization was achieved by double-thymidine treatment. The expression levels of *Hox* genes were determined by semi-quantitative RT–PCR[6]. The pattern of Ubp-M, Bmi1 and ubH2A binding to the *HOXD10* gene was determined by ChIP assays. Effects of Ubp-M on body transformation were determined by injection of IgG and Ubp-M antibody into *Xenopus laevis* embryos at the two-cell stage. Expression of the *Hoxd10* gene and posterior and anterior molecular markers was revealed by *in situ* hybridization.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Martin, C. & Zhang, Y. The diverse functions of histone lysine methylation. *Nature Rev. Mol. Cell Biol.* **6**, 838–849 (2005).
2. Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293**, 1074–1080 (2001).
3. Peterson, C. L. & Laniel, M.-A. Histones and histone modifications. *Curr. Biol.* **14**, R546–R551 (2004).
4. Wang, H. *et al.* Role of histone H2A ubiquitination in Polycomb silencing. *Nature* **431**, 873–878 (2004).
5. Cao, R., Tsukada, Y.-i. & Zhang, Y. Role of Bmi-1 and Ring1A in H2A ubiquitylation and *Hox* gene silencing. *Mol. Cell* **20**, 845–854 (2005).
6. Wei, J., Zhai, L., Xu, J. & Wang, H. Role of Bmi1 in H2A ubiquitylation and *Hox* gene silencing. *J. Biol. Chem.* **281**, 22537–22544 (2006).
7. Fang, J., Chen, T., Chadwick, B., Li, E. & Zhang, Y. Ring1b-mediated H2A ubiquitination associates with inactive X chromosomes and is involved in initiation of X inactivation. *J. Biol. Chem.* **279**, 52812–52815 (2004).
8. de Napoles, M. *et al.* Polycomb group proteins Ring1A/B link ubiquitylation of histone H2A to heritable gene silencing and X inactivation. *Dev. Cell* **7**, 663–676 (2004).
9. Fang, J., Wang, H. & Zhang, Y. Purification of histone methyltransferases from HeLa cells. *Methods Enzymol.* **377**, 213–226 (2004).
10. Wang, H. *et al.* Purification and functional characterization of a histone H3-lysine 4-specific methyltransferase. *Mol. Cell* **8**, 1207–1217 (2001).
11. Henry, K. W. *et al.* Transcriptional activation via sequential histone H2B ubiquitylation and deubiquitylation, mediated by SAGA-associated Ubp8. *Genes Dev.* **17**, 2648–2663 (2003).
12. Daniel, J. A. *et al.* Deubiquitination of histone H2B by a yeast acetyltransferase complex regulates transcription. *J. Biol. Chem.* **279**, 1867–1871 (2004).
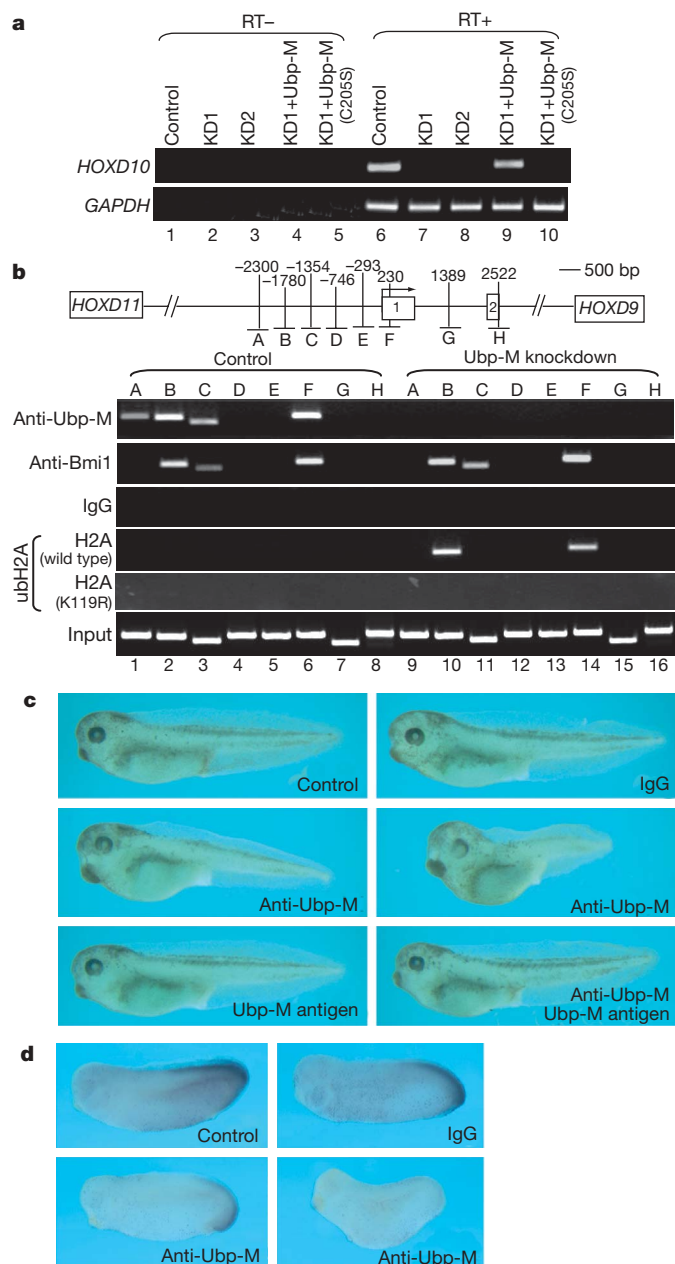
13. Erdjument-Bromage, H. *et al.* Examination of micro-tip reversed-phase liquid chromatographic extraction of peptide pools for mass spectrometric analysis. *J. Chromatogr. A.* **826,** 167–181 (1988).

14. Winkler, G. *et al.* Isolation and mass spectrometry of transcription factor complexes. *Methods* **26,** 260–269 (2002).

15. Cai, S. Y., Babbitt, R. & Marchesi, V. A mutant deubiquitinating enzyme (Ubp-M) associates with mitotic chromosomes and blocks cell division. *Proc. Natl Acad. Sci. USA* **96,** 2828–2833 (1999).

16. Mimnaugh, E. G. *et al.* Caspase-dependent deubiquitination of monoubiquitinated nucleosomal histone H2A induced by diverse apoptogenic stimuli. *Cell Death Differ.* **8,** 1182–1196 (2001).

17. Wang, H. *et al.* Histone H3 and H4 ubiquitylation by the CUL4-DDB-ROC1 ubiquitin ligase facilitates cellular response to DNA damage. *Mol. Cell* **22,** 383–394 (2006).

18. Shiio, Y. & Eisenman, R. N. Histone sumoylation is associated with transcriptional repression. *Proc. Natl Acad. Sci. USA* **100,** 13225–13230 (2003).

19. Robzyk, K., Recht, J. & Osley, M. A. Rad6-dependent ubiquitination of histone H2B in yeast. *Science* **287,** 501–504 (2000).

Author Contributions H.W. designed the experimental strategy, performed parts of the purification and wrote the paper; H.-Y.J. performed most of the purification, determined the substrate preference and investigated the role of Ubp-M in cell cycle progression and gene expression; L.Z. determined the substrate specificity of Ubp-M and performed the *in vitro* kinase reaction and nucleosome pull-down experiments; C.Y. cloned the *Xenopus Hoxd10* gene and generated the RNAi-resistant Ubp-M constructs; S.N. and C.C. performed the *Xenopus* injection and *in situ* hybridization; H.E.-B. and P.T. performed mass spectrometric analysis.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to H.W. (hbwang@uab.edu).

METHODS

Substrate preparation and *in vitro* histone deubiquitination assay. H2A-containing mononucleosomes were purified by tandem immunoprecipitation[6]. Histone octamers, mononucleosomes and oligonucleosomes containing ubH2A were prepared as described[10]. Flag–H2A was purified from HeLa stable cell lines and Flag–H2B was purified from the T85 yeast strain and the YKH045 yeast strain with anti-Flag immunoprecipitation under denaturing conditions[17,20,21]. Recombinant H3, H4, H2A and H2B were purified from *Escherichia coli*[22]. To purify individual H2A, ubH2A, H2B and ubH2B, gel strips containing the corresponding histones were excised and proteins were electro-eluted out and further passed through a hydroxyapatite column to remove the excess SDS. Histone octamer refolding and nucleosome reconstitution were performed as described[17,23].

Histone deubiquitination reactions were performed as follows. 1.5 μg of ubH2A-containing mononucleosomes, histone octamers, oligonucleosomes, or 0.8 μg of reconstituted nucleosomes were incubated with protein fractions or recombinant Ubp-M in deubiquitination reaction buffer (100 mM Tris-HCl, pH 8.0, 1 mM EDTA, 0.1 mM PMSF, 1 mM dithiothreitol (DTT), 1 μg ml^{-1} aprotinin, 1 μg ml^{-1} leupeptin and 1 μg ml^{-1} pepstatin A) at 37 °C for 45 min, or as indicated. The reaction was terminated by the addition of SDS–PAGE sample loading buffer and proteins were resolved on SDS–PAGE and blotted with the anti-HA antibody.

Purification of native and recombinant Ubp-M. HeLa nuclear proteins were separated into nuclear extracts and nuclear pellets using a previously described procedure[24]. Nuclear extract (6 g) was loaded onto a 700 ml P11 column equilibrated with buffer C (20 mM Tris-HCl, pH 7.9, 0.2 mM EDTA, 1 mM DTT, 0.1 mM PMSF, 0.025% NP-40, 10% glycerol) containing 100 mM KCl (BC100). Proteins that bound to the column were step eluted with BC300, BC500 and BC1000. The BC500 fraction was dialysed against buffer D (40 mM HEPES-KOH, pH 7.9, 0.2 mM EDTA, 1 mM DTT, 0.1 mM PMSF, 0.025% NP-40, 10% glycerol) containing 20 mM ammonium sulphate (BD20) and then loaded onto a 45 ml HPLC-DEAE-5PW column (TOSOH Bioscience). Bound proteins were eluted with an 8-column-volume linear gradient from BD20 to BD500. Fractions that contain H2A-deubiquitination activity were adjusted to BD600 with saturated ammonium sulphate and loaded onto a 22 ml FPLC phenyl sepharose column (Amersham Biosciences). Bound proteins were eluted with a 12-column-volume linear gradient from BD600 to BD0. The active fractions were then applied to a Sephacryl S-300 column (Amersham Biosciences). The H2A-deubiquitination activity eluted between 443–670 kDa. These fractions were dialysed against buffer P (5 mM HEPES-KOH, pH 7.5, 0.04 M KCl, 0.01 mM CaCl$_2$, 10% glycerol, 1 mM DTT, 0.1 mM PMSF) containing 10 mM potassium phosphate and loaded onto a 1 ml hydroxyapatite column (Bio-rad). The column was eluted with a 20-column-volume linear gradient from BP10 to BP600. Fractions that contained H2A-deubiquitination activity were loaded onto a Superose 6 column (Amersham Biosciences) that was equilibrated with BC500. To identify the 120-kDa protein that co-elutes with the H2A deubiquitination activity, fractions 58–60 of the Superose 6 column were concentrated and resolved on SDS–PAGE. The band that contains the 120-kDa protein was excised and subjected to in-gel trypsin digestion. The resulting peptides were analysed by matrix-assisted laser desorption/ionization reflectron time-of-flight (MALDI-reTOF) mass spectrometry (MS) (UltraFlex TOF/TOF; BRUKER) for peptide mass fingerprinting[13,14]. Recombinant Ubp-M was purified from Sf9 cells infected with baculovirus encoding Flag–Ubp-M as described[6].

Ubp-M knockdown, RT–PCR and ChIP assays. Stable cell lines with knockdown of Ubp-M were generated based on a vector-based knockdown strategy[4,17]. For transient knockdown, siRNA oligonucleotides against Ubp-*M* were purchased from Invitrogen in a purified, annealed duplex form and transfected into cells with Lipofectamine 2000[6,17]. Semi-quantitative RT–PCR was performed as described[6].

ChIP and double ChIP were performed as described except that cells were lysed in buffer containing 1% SDS and diluted 10 times before immunoprecipitation[6]. All PCR reactions were performed within the range of linear amplification. The sequences for RT–PCR and ChIP assays are available on request.

The DNA sequence encoding the stem hairpin RNA for Ubp-M was 5′-ACCAGTGCTTAGAGAACTATCTCTTGAATAGTTCTCTAAGCACTGGT-3′ (KD1) and 5′-ACCAGTGCTTAGAGAACTATTCAAGAGATAGTTCTCTAA-GCACTGGT-3′ (KD2). Underlined letters represent targeting sequences of Ubp-M. For transient knockdown, siRNA oligonucleotides against Ubp-M were purchased from Invitrogen in a purified, annealed duplex form. The sequences for these siRNA are as follows: siRNA1 5′-CCUCCUGUUCUUACUCUU-CAUUUAA-3′ and 5′-UUAAAUGAAGAGUAAGAACAGGAGG-3′; siRNA2 5′-CCGGAAAUCUUAGAUUUGGCUCCUU-3′ and 5′-AAGGAGCCAAAUC-UAAGAUUUCCGG-3′. siRNA1 and siRNA2 were mixed at a 1:1 ratio and transfected into cells.

Cell cycle and FACS analysis. Cells were synchronized by double thymidine (Sigma) treatment[25]. After release into fresh medium, cells were harvested at the indicated time points. Cells were fixed in 70% ethanol, permeabilized with 0.25% Triton X-100 and subsequently incubated with anti-phospho-histone H3 antibody (Ser 10; Upstate), APC-conjugated anti-rabbit antibody (Invitrogen) and propidium iodide/RNase A. Mitotic cell distribution was analysed by FACS. Alternatively, acid-soluble proteins were extracted from control and Ubp-M knockdown cells and levels of ubH2A and H3S10ph were analysed by western blot assays.

***Xenopus* manipulations.** A total of 300–600 pg IgG or anti-Ubp-M antibody was injected into both blastomeres of two-cell-stage embryos. The embryos were collected at tailbud stages for *in situ* hybridization with *Hoxd10, Sox2, Otx2, En2* and *Krox20* antisense RNA probes, or at tadpole stages for morphological assessment of the effects. For specificity control, an equal amount of Ubp-M antigen was mixed with the antibody on ice for at least 30 min before they were co-injected into the embryos. Whole-mount *in situ* hybridization was performed as described[26].

Constructs, antibodies, kinase reactions and immunofluorescence staining. The cDNA for Ubp-M was provided by V. T. Marchesi[15]. To generate antibody against Ubp-M, the amino terminus of Ubp-M (1–257 amino acids) was cloned into pGEX-KG and the purified recombinant proteins were used to immunize rabbits. To construct RNAi-resistant Ubp-M constructs, a point mutation was introduced into siRNA targeting nucleotide acid sequence but without changing the encoded amino acid. Site-directed mutagenesis was performed as described[4]. For baculovirus production, Ubp-M cDNA was cloned into pFastBacHTb vector (Gibco) with a Flag tag at its N terminus. Recombinant baculovirus expressing Flag–Ubp-M was generated and amplified following the manufacturer's protocol. Kinase assay was performed as described[27]. Aurora B kinase was purified by immunoprecipitation from HeLa cells that stably express Flag–Aurora B and the phosphorylation of H3 was determined by western blot with anti-phospho-H3S10 antibody (Upstate, 06-570). Immunofluorescence staining was performed essentially as described[28,29].

20. Henry, K. W. *et al.* Transcriptional activation via sequential histone H2B ubiquitylation and deubiquitylation, mediated by SAGA-associated Ubp8. *Genes Dev.* **17,** 2648–2663 (2003).

21. Emre, N. C. T. *et al.* Maintenance of low histone ubiquitylation by Ubp10 correlates with telomere-proximal sir2 association and gene silencing. *Mol. Cell* **17,** 585–594 (2005).

22. Luger, K., Rechsteiner, T. J. & Richmond, T. J. Preparation of nucleosome core particle from recombinant histones. *Methods Enzymol.* **304,** 3–19 (1999).

23. Dyer, P. N. *et al.* Reconstitution of nucleosome core particles from recombinant histones and DNA. *Methods Enzymol.* **375,** 23–44 (2004).

24. Dignam, J. D., Martin, P., Shastry, B. & Roeder, R. Eukaryotic gene transcription with purified components. *Methods Enzymol.* **101,** 582–598 (1983).

25. Fischle, W. *et al.* Regulation of HP1-chromatin binding by histone H3 methylation and phosphorylation. *Nature* **438,** 1116–1122 (2005).

26. Harland, R. M. *In situ* hybridization: an improved whole-mount method for *Xenopus* embryos. *Methods Cell Biol.* **36,** 685–695 (1991).

27. Bolton, M. A. *et al.* Aurora B kinase exists in a complex with survivin and INCENP and its kinase activity is stimulated by survivin binding and phosphorylation. *Mol. Biol. Cell* **13,** 3064–3077 (2002).

28. Perez-Burgos, L. *et al.* Generation and characterization of methyl-lysine histone antibodies. *Methods Enzymol.* **376,** 234–254 (2003).

29. Song, L. *et al.* Ser-10 phosphorylated histone H3 is involved in cytokinesis as a chromosomal passenger. *Cell Biol. Int.* **31,** 1184–1190 (2007).

LETTERS

An elaborate pathway required for Ras-mediated epigenetic silencing

Claude Gazin[1]*, Narendra Wajapeyee[1]*, Stephane Gobeil[1], Ching-Man Virbasius[1] & Michael R. Green[1]

The conversion of a normal cell to a cancer cell occurs in several steps and typically involves the activation of oncogenes and the inactivation of tumour suppressor and pro-apoptotic genes[1]. In many instances, inactivation of genes critical for cancer development occurs by epigenetic silencing, often involving hypermethylation of CpG-rich promoter regions[2,3]. It remains to be determined whether silencing occurs by random acquisition of epigenetic marks that confer a selective growth advantage or through a specific pathway initiated by an oncogene[4–6]. Here we perform a genome-wide RNA interference (RNAi) screen in *K-ras*-transformed NIH 3T3 cells and identify 28 genes required for Ras-mediated epigenetic silencing of the pro-apoptotic *Fas* gene. At least nine of these RESEs (Ras epigenetic silencing effectors), including the DNA methyltransferase DNMT1, are directly associated with specific regions of the *Fas* promoter in *K-ras*-transformed NIH 3T3 cells but not in untransformed NIH 3T3 cells. RNAi-mediated knockdown of any of the 28 RESEs results in failure to recruit DNMT1 to the *Fas* promoter, loss of *Fas* promoter hypermethylation, and derepression of Fas expression. Analysis of five other epigenetically repressed genes indicates that Ras directs the silencing of multiple unrelated genes through a largely common pathway. Last, we show that nine RESEs are required for anchorage-independent growth and tumorigenicity of *K-ras*-transformed NIH 3T3 cells; these nine genes have not previously been implicated in transformation by Ras. Our results show that Ras-mediated epigenetic silencing occurs through a specific, complex, pathway involving components that are required for maintenance of a fully transformed phenotype.

Members of the *ras* oncogene family transform most immortalized cell lines, and mutations of *ras* genes occur in about 30% of human tumours[7]. In addition, activation of the Ras pathway is frequent in human tumours even in the absence of *ras* mutations[8]. Previous studies have shown that, in mouse NIH 3T3 cells, activated Ras epigenetically silences *Fas* expression, thereby preventing Fas-ligand-induced apoptosis[9,10]. Activated Ras also epigenetically silences *Fas* expression in the human *K-ras*-transformed cell line, HEC1A (Supplementary Fig. 1). To identify genes required for Ras-mediated silencing of *Fas*, we performed a genome-wide small hairpin RNA (shRNA) screen using, as a selection strategy, re-expression of Fas protein on the cell surface (Fig. 1a). A mouse shRNA library comprising about 62,400 shRNAs directed against about 28,000 genes was divided into 10 pools, which were packaged into retrovirus particles and used in the stable transduction of Fas-negative *K-ras* NIH 3T3 cells. Fas-positive cells in each pool were selected on immunomagnetic beads by using an anti-Fas antibody; the Fas-positive population was expanded and the shRNAs were identified by sequence analysis. Positive candidates were confirmed by stably transducing

K-ras NIH 3T3 cells with single shRNAs directed against the candidate genes, followed by immunoblot analysis for Fas re-expression.

The screen identified 28 genes that, after shRNA-mediated knockdown, resulted in Fas re-expression. These genes are listed in Table 1, and an immunoblot analysis of Fas re-expression in the 28 *K-ras* NIH 3T3 knockdown (*K-ras* NIH 3T3 KD) cell lines is shown in Fig. 1b. Consistent with previous reports[10], treatment of *K-ras* NIH 3T3 cells with the DNA methylation inhibitor 5-aza-2′-deoxycytidine (5-aza) restored Fas expression (see also Supplementary Fig. 1). Quantitative

Figure 1 | A genome-wide shRNA screen identifies factors required for Ras-mediated epigenetic silencing of Fas. a, Schematic summary of the genome-wide shRNA screen for Ras-mediated epigenetic silencing of Fas. **b**, Immunoblot analysis monitoring Fas expression in the 28 *K-ras* NIH 3T3 KD cell lines. Expression of Fas in *K-ras* NIH 3T3 cells in the presence and absence of 5-aza is also shown. *K-*Ras expression is shown as a loading control.

[1]Howard Hughes Medical Institute and Programs in Gene Function and Expression and Molecular Medicine, University of Massachusetts Medical School, 364 Plantation Street, Worcester, Massachusetts 01605, USA.
*These authors contributed equally to this work.

Table 1 | Genes required for Ras-mediated epigenetic silencing of *Fas*

Biological process	Accession number	Gene symbol	Name
Signal transduction	XM_993034	*Kalrn*	Kalirin, RhoGEF kinase
	NM_011949	*Mapk1*	Mitogen-activated protein kinase 1
	NM_177395	*Map3k9*	Mitogen-activated protein kinase kinase kinase 9
	NM_011062	*Pdpk1*	3-Phosphoinositide-dependent protein kinase 1
	NM_172498	*Ptk2b*	PTK2 protein tyrosine kinase 2 beta
	XM_193738	*S100z*	S100 calcium-binding protein, zeta
Transcription regulation	NM_181322	*Ctcf*	CCCTC-binding factor
	NM_025613	Eid1	EP300 interacting inhibitor of differentiation 1
	NM_007891	*E2f1*	E2F transcription factor 1
	NM_054048	*Rcor2*	REST co-repressor 2
	XM_284529	*Sox14*	SRY-box containing gene 14
	NM_181853	*Trim66*	Tripartite motif-containing protein 66
	NM_013744	*Zfp354b*	Zinc finger protein 354B
Chromatin modification	NM_007552	*Bmi1*	Bmi1 polycomb ring finger oncogene
	NM_010066	*Dnmt1*	DNA methyltransferase (cytosine-5) 1
	NM_199322	*Dot1l*	DOT1-like, histone H3 methyltransferase (*Saccharomyces cerevisiae*)
	NM_021876	*Eed*	Embryonic ectoderm development
	NM_007971	*Ezh2*	Enhancer of zeste homologue 2 (*Drosophila*)
	NM_024124	*Hdac9*	Histone deacetylase 9
	NM_028479	*Mrgbp*	MRG-binding protein
	NM_009762	*Smyd1*	SET and MYND domain containing 1
Chromatin remodelling	NM_025541	*Asf1a*	ASF1 anti-silencing function 1 homologue A (*S. cerevisiae*)
	NM_054078	*Baz2a*	Bromodomain adjacent to zinc finger domain, 2A
	NM_181345	*Npm2*	Nucleophosmin/nucleoplasmin 2
Genome stability/ageing	NM_181586	*Sirt6*	Sirtuin 6 (silent mating type information regulation 2, homologue) 6 (*S. cerevisiae*)
Unknown	XM_146572	*Sipa1l2*	Signal-induced proliferation-associated gene 1 like 2
	NM_197987	*Trim37*	Tripartite motif-containing protein 37
	XM_132052	*Zcchc4*	Zinc finger, CCHC domain containing 4

real-time reverse-transcriptase-mediated polymerase chain reaction (qRT–PCR) confirmed in all cases that expression of the target gene was decreased in each *K-ras* NIH 3T3 KD cell line (Supplementary Fig. 2). For all 28 genes a second, unrelated, shRNA directed against the same target also resulted in *Fas* re-expression when stably expressed in *K-ras* NIH 3T3 cells (Supplementary Fig. 3). Knockdown of each of these 28 genes in an additional cell line, *H-ras*-transformed murine C3H10T1/2 cells, also derepressed the epigenetically silenced *Fas* gene (Supplementary Fig. 4a).

For convenience we refer to the protein products of the 28 genes as RESEs. The RESEs include cytoplasmic cell signalling molecules and nuclear regulators of gene expression (Table 1). Among the cell signalling components, PDPK1, a serine/threonine kinase, is known to function downstream of Ras and to regulate the phosphatidylinositol-3-OH kinase (PI(3)K)–Akt pathway, which is frequently activated in cancer[11]. Significantly, it has been previously reported that the PI(3)K–Akt pathway is involved in Ras-mediated silencing of *Fas*[10]. Other cell signalling proteins include MAPK1, a proximal Ras target that is frequently activated in cancer[12], and PTK2B, which is recruited to cell membranes by activated Ras[13].

Among the nuclear gene regulatory proteins are known transcriptional activators and repressors/co-repressors (CTCF, EID1, E2F1, RCOR2 and TRIM66/TIF1D) including several predicted sequence-specific DNA-binding proteins (SOX14, ZCCHC4 and ZFP345B), three histone methyltransferases (DOT1L, EZH2 and SMYD1), a histone deacetylase (HDCA9), two histone chaperones (ASF1A and NPM2), the maintenance DNA methyltransferase DNMT1, and several Polycomb group proteins (BMI1, EED and EZH2). Several recent studies have linked Polycomb proteins to abnormal DNA methylation and gene silencing[14–16]. One of the nuclear RESEs is BAZ2A/TIP5, previously known only to be involved in the repression of RNA polymerase I-directed ribosomal gene transcription[17]. Several RESEs were substantially upregulated at the transcriptional (Supplementary Fig. 5) or post-transcriptional (Supplementary Fig. 6) level in *K-ras* NIH 3T3 cells compared with NIH 3T3 cells, which explains, at least in part, how *K-ras* activates this silencing pathway.

As mentioned above, treatment of *K-ras* NIH 3T3 cells with 5-aza results in *Fas* re-expression, suggesting that repression is due, at least in part, to promoter hypermethylation. We therefore sought to

determine the relationship between *Fas* promoter hypermethylation and *Fas* re-expression after knockdown of each of the 28 RESEs. We first confirmed that the repressed *Fas* promoter was hypermethylated and mapped the hypermethylated region(s) by bisulphite sequence analysis. These results, summarized in Fig. 2a, reveal three regions located upstream and downstream from the transcription start site (TSS) that are hypermethylated in *K-ras* NIH 3T3 cells but not in NIH 3T3 cells or in *K-ras* NIH 3T3 cells after knockdown of DNMT1. Significantly, these same three *Fas* promoter regions are also hypermethylated in *H-ras*-transformed C3H10T1/2 cells but not in C3H10T1/2 cells or in *H-ras*-transformed C3H10T1/2 cells after knockdown of DNMT1 (Supplementary Fig. 4b).

To facilitate analysis of the methylation status of these three regions in the 28 *K-ras* NIH 3T3 KD cell lines we established and validated a rapid methylated DNA immunoprecipitation (MeDIP) assay, in which the antibody is directed against 5-methylcytosine[18]. The MeDIP results of Supplementary Fig. 7 show that in *K-ras* NIH 3T3 cells the *Fas* promoter was hypermethylated within the TSS/downstream region, which is consistent with the bisulphite sequencing results. Moreover, the MeDIP results show, as expected, that the TSS/downstream region was not hypermethylated in NIH 3T3 cells or in *K-ras* NIH 3T3 cells after treatment with 5-aza. We then assessed the three hypermethylated *Fas* promoter regions in each of the 28 *K-ras* NIH 3T3 KD cell lines. The results of Fig. 2b show that in all 28 *K-ras* NIH 3T3 KD cell lines the three *Fas* promoter regions were not hypermethylated, which is consistent with the expression data.

For a better understanding of the basis of *Fas* silencing, we examined whether nuclear RESEs functioned by direct association with the *Fas* promoter. We performed a series of chromatin immunoprecipitation (ChIP) assays, based on antibody availability, using three sets of promoter-specific primer pairs located about 2 kilobases (kb) upstream of the TSS, about 1 kb upstream of the TSS or encompassing the core promoter/TSS. The three primer pairs cover the entire *Fas* promoter region. Figure 2c shows that in *K-ras* NIH 3T3 cells, nine of the RESEs were bound to specific *Fas* promoter regions: NPM2, TRIM66 and ZFP354B were present about 2 kb upstream of the TSS; BMI1, DNMT1, SIRT6 and TRIM37 were present about 1 kb upstream of the TSS; and EZH2, CTCF and NPM2 were present

at the core promoter/TSS. In NIH 3T3 cells only NPM2 was detectably associated with the *Fas* promoter at the core promoter/TSS. Thus, as summarized in Fig. 2d, at least nine RESEs are recruited to specific regions of the *Fas* promoter in response to expression of activated Ras.

The ChIP results of Fig. 2c show that DNMT1 is associated with the *Fas* promoter in *K-ras* NIH 3T3 cells but not in untransformed NIH 3T3 cells. The two other DNA methyltransferases, DNMT3A and DNMT3B, were not identified in the original shRNA screen and are not detectably associated with the *Fas* promoter by ChIP analysis (Supplementary Fig. 8). These results strongly suggest that

Figure 2 | ChIP analysis and methylation status of the *Fas* promoter.
a, Summary of bisulphite sequencing. Open circles, unmethylated CpG; filled circles, methylated CpG. Each row represents a single clone. CpG dinucleotide positions are shown by vertical lines. **b**, MeDIP analysis after knockdown of each of the 28 RESEs. NS, nonsilencing shRNA. Values are expressed as the fold difference relative to input, and have been corrected for background. Error bars indicate s.e.m. (*n* = 3). **c**, ChIP assay monitoring occupancy of selected RESEs at the core promoter/TSS (CP/TSS) or about 1 kb or about 2 kb upstream of the TSS. **d**, Summary of the ChIP results. **e**, ChIP analysis monitoring occupancy of DNMT1. Error bars indicate s.e.m. (*n* = 3).

DNMT1 is required to sustain hypermethylation of the *Fas* promoter in *K-ras* NIH 3T3 cells. To confirm this possibility, we analysed the association of DNMT1 with the *Fas* promoter in the 28 *K-ras* NIH 3T3 KD cell lines. The ChIP results of Fig. 2e show that in all 28 *K-ras* NIH 3T3 KD cell lines, association of DNMT1 with the *Fas* promoter was markedly decreased. Moreover, bisulphite sequence analysis showed that after knockdown of DNMT1 the TSS/downstream region of the *Fas* promoter was no longer hypermethylated (see Fig. 2a).

Several genes in addition to *Fas* are known to be epigenetically silenced in *ras*-transformed cells. To gain insight into whether Ras mediates epigenetic silencing of different genes through common or diverse pathways, we analysed five other well-studied, epigenetically silenced genes: *Sfrp1*, *Par4*/*Pawr*, *Plagl1*, *H2-K1* and *Lox*. A variety of evidence supports the relevance of these genes to cellular transformation and cancer (reviewed in refs 19–23). The results of Fig. 3a show that, like *Fas*, all five genes were expressed in NIH 3T3 cells but not in *K-ras* NIH 3T3 cells, and were re-expressed in *K-ras* NIH 3T3 cells after treatment with 5-aza. Bisulphite sequence analysis confirmed that all five genes contained regions that are hypermethylated in *K-ras* NIH 3T3 cells but not in NIH 3T3 cells or in *K-ras* NIH 3T3 cells after knockdown of DNMT1 (Supplementary Fig. 9). For four of these genes (*Sfrp1*, *Par4*, *Plagl1* and *H2-K1*), the TSS was encompassed by dense hypermethylation in *K-ras* NIH 3T3 cells.

We next analysed the expression of *Sfrp1*, *Par4*, *Plagl1*, *H2-K1* and *Lox* in the 28 *K-ras* NIH 3T3 KD cell lines. The qRT–PCR results of Supplementary Fig. 10 are summarized in Fig. 3b and reveal substantial overlap in the requirements of RESEs for epigenetic silencing of *Fas*, *Sfrp1*, *Par4*, *Plagl1*, *H2-K1* and *Lox*: of the 28 RESEs required for silencing of *Fas*, at least 21 were also required for silencing of each of the five other genes analysed. MeDIP analysis for all five genes revealed a perfect correspondence between the RESEs required for silencing and for promoter hypermethylation (Supplementary Fig. 11). These results indicate that Ras directs the epigenetic silencing of multiple, unrelated genes through a largely common pathway.

Proteins that function downstream of Ras could be essential for a fully transformed phenotype. To determine whether any of the 28 RESEs were also required for Ras-mediated transformation, we first tested the ability of the *K-ras* NIH 3T3 KD cell lines to grow in soft agar. Supplementary Fig. 12a shows that knockdown of any of nine RESEs (S100Z, MRGBP, BAZ2A, SMYD1, EID1, TRIM66, TRIM37, ZCCHC4 and KALRN) markedly inhibited anchorage-independent growth.

For a further characterization of the role of these nine RESEs in Ras-mediated transformation, we tested the ability of the corresponding nine *K-ras* NIH 3T3 KD cell lines to form tumours after subcutaneous injection in the flanks of nude mice. Supplementary Fig. 12b shows that knockdown of SMYD1 or BAZ2A moderately inhibited tumour growth, whereas knockdown of S100Z, TRIM37, TRIM66, EID1, ZCCHC4, MRGBP or KALRN markedly inhibited tumour growth.

It is well established that in many cancers specific genes affecting cellular growth control are hypermethylated and epigenetically silenced[2,3]. However, the mechanistic basis of epigenetic silencing is not understood. The results presented here show that oncogenic Ras directs epigenetic silencing through a specific, complex pathway involving at least 28 components (RESEs). Knockdown of any of the 28 RESEs resulted in a failure to recruit DNMT1 to the *Fas* promoter, loss of *Fas* promoter hypermethylation, and derepression of Fas expression. The vast majority of RESEs have not been previously connected to the Ras pathway; our results have therefore identified several new factors that act downstream of Ras. Nine RESEs are required for anchorage-independent growth and tumorigenicity. The identification of new components that act downstream of Ras, and are required for epigenetic silencing and complete transformation, provides potential new anti-cancer targets.
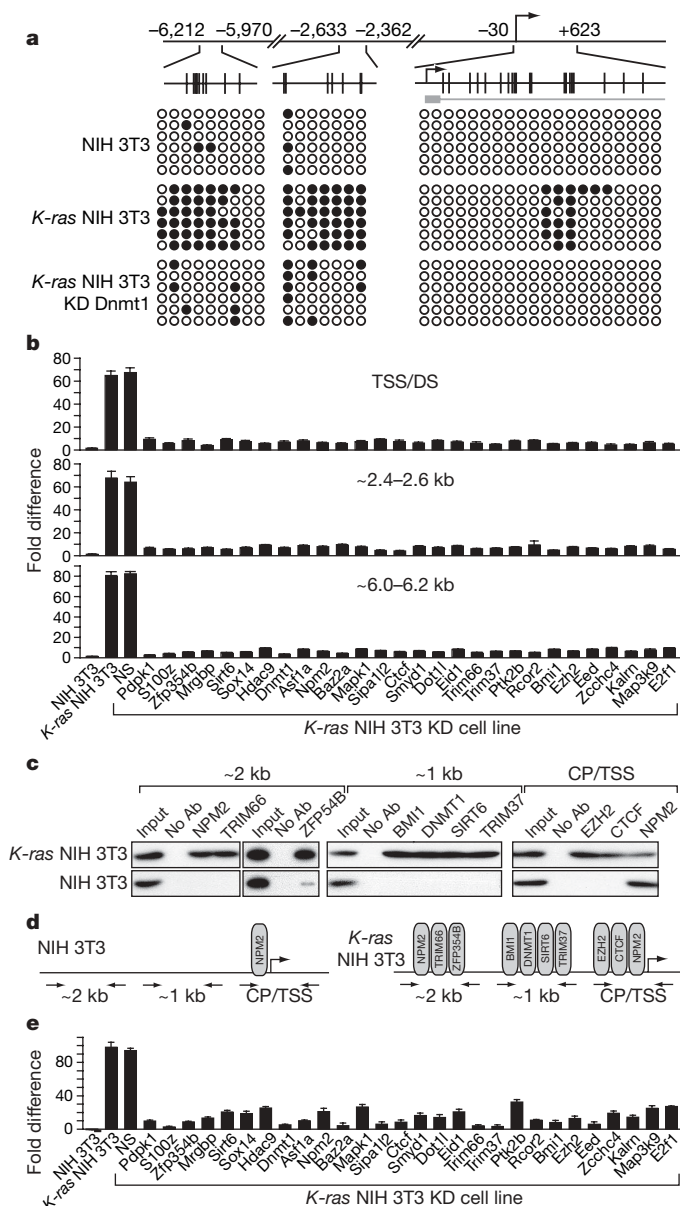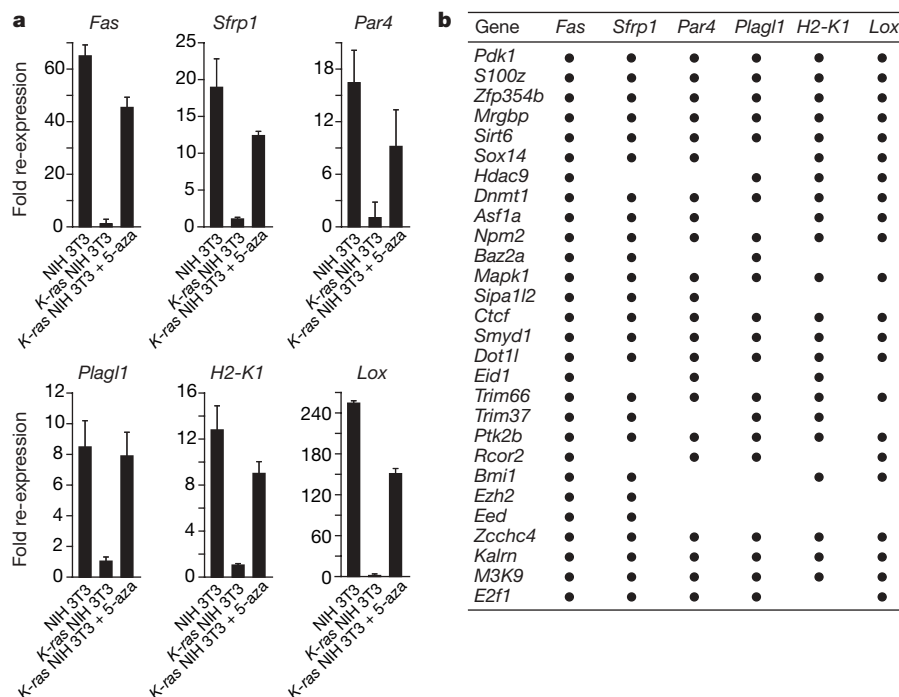
Figure 3 | Ras directs epigenetic silencing of multiple, unrelated genes through a largely common pathway. a, qRT–PCR monitoring expression of *Fas*, *Sfrp1*, *Par4*, *Plagl1*, *H2-K1* and *Lox* in NIH 3T3 cells, and in *K-ras* NIH 3T3 cells in the presence and absence of 5-aza. Values are expressed as fold re-expression relative to expression of the gene in *K-ras* NIH 3T3 cells, which is arbitrarily set to unity. Error bars indicate s.e.m. ($n = 3$). **b**, Summary of qRT–PCR analysis monitoring the re-expression of *Fas*, *Sfrp1*, *Par4*, *Plagl1*, *H2-K1* and *Lox* after knockdown of each of the 28 RESEs.

METHODS SUMMARY

shRNA screen. The mouse shRNA[mir] library (release 2.16; Open Biosystems) was used to generate ten retroviral pools, each comprising about 6,000 shRNA clones. *K-ras* NIH 3T3 cells were transduced with the retroviral pools, and two days later they were selected for seven days for resistance to puromycin. Fas-positive cells were selected and expanded, and genomic DNA was isolated. To identify the candidate shRNAs, the shRNA region of the transduced virus was amplified by PCR, then cloned and sequenced. Individual knockdown cell lines were generated by retroviral transduction of *K-ras* NIH 3T3 cells with the respective shRNA. Individual shRNAs were either obtained from the Open Biosystems library or synthesized (see Supplementary Table 1).

Bisulphite sequencing. Bisulphite modification was conducted essentially as described[24] except that hydroquinone was used at a concentration of 125 mM during bisulphite treatment performed in the dark, and DNA was desalted on Qiaquick columns (Qiagen) after the bisulphite reaction.

Chromatin immunoprecipitation (ChIP) and methylated DNA immuno-precipitation (MeDIP). ChIP assays were performed on extracts prepared seven days after retroviral transduction and selection with puromycin. The sequences of the primers used for amplifying the MeDIP and ChIP products are shown in Supplementary Table 2.

Quantitative real-time RT–PCR. Total RNA was isolated and reverse transcription was performed, followed by qRT–PCR. Primer sequences are provided in Supplementary Table 2.

Soft agar assays. Soft agar assays were performed with the CytoSelect 96-well Cell Transformation Assay (Cell Biolabs) in accordance with the manufacturer's instructions.

Tumour formation assays. NIH 3T3, *K-ras* NIH 3T3, or *K-ras* NIH 3T3 knock-down cells (5×10^6) were injected subcutaneously into the right flank of athymic Balb/c (*nu/nu*) mice (Taconic). Tumour dimensions were measured every three days from the time of appearance of the tumours. Animal experiments were performed in accordance with the Institutional Animal Care and Use Committee (IACUC) guidelines.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 23 July; accepted 7 September 2007.

1. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
2. Baylin, S. B. DNA methylation and gene silencing in cancer. *Nature Clin. Pract. Oncol.* **2** (Suppl. 1), S4–S11 (2005).
3. Esteller, M. Epigenetics provides a new generation of oncogenes and tumour-suppressor genes. *Br. J. Cancer* **94**, 179–183 (2006).
4. Jones, P. A. DNA methylation errors and cancer. *Cancer Res.* **56**, 2463–2467 (1996).
5. Baylin, S. & Bestor, T. H. Altered methylation patterns in cancer cell genomes: cause or consequence? *Cancer Cell* **1**, 299–305 (2002).
6. Keshet, I. *et al.* Evidence for an instructive mechanism of *de novo* methylation in cancer cells. *Nature Genet.* **38**, 149–153 (2006).
7. Giehl, K. Oncogenic Ras in tumour progression and metastasis. *Biol. Chem.* **386**, 193–205 (2005).
8. Ehmann, F. *et al.* Detection of N-RAS and K-RAS in their active GTP-bound form in acute myeloid leukemia without activating RAS mutations. *Leuk. Lymphoma* **47**, 1387–1391 (2006).
9. Fenton, R. G., Hixon, J. A., Wright, P. W., Brooks, A. D. & Sayers, T. J. Inhibition of Fas (CD95) expression and Fas-mediated apoptosis by oncogenic Ras. *Cancer Res.* **58**, 3391–3400 (1998).
10. Peli, J. *et al.* Oncogenic Ras inhibits Fas ligand-mediated apoptosis by downregulating the expression of Fas. *EMBO J.* **18**, 1824–1831 (1999).
11. Osaki, M., Oshimura, M. & Ito, H. PI3K-Akt pathway: its functions and alterations in human cancer. *Apoptosis* **9**, 667–676 (2004).
12. de Vries-Smits, A. M., Burgering, B. M., Leevers, S. J., Marshall, C. J. & Bos, J. L. Involvement of p21^ras in activation of extracellular signal-regulated kinase 2. *Nature* **357**, 602–604 (1992).
13. Alfonso, P. *et al.* Proteomic analysis of p38α mitogen-activated protein kinase-regulated changes in membrane fractions of RAS-transformed fibroblasts. *Proteomics* **6** (Suppl 1), S262–S271 (2006).
14. Ohm, J. E. *et al.* A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nature Genet.* **39**, 237–242 (2007).
15. Schlesinger, Y. *et al.* Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for *de novo* methylation in cancer. *Nature Genet.* **39**, 232–236 (2007).
16. Widschwendter, M. *et al.* Epigenetic stem cell signature in cancer. *Nature Genet.* **39**, 157–158 (2007).
17. Zhou, Y., Santoro, R. & Grummt, I. The chromatin remodeling complex NoRC targets HDAC1 to the ribosomal gene promoter and represses RNA polymerase I transcription. *EMBO J.* **21**, 4632–4640 (2002).
18. Weber, M. *et al.* Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genet.* **37**, 853–862 (2005).

19. Rubin, J. S., Barshishat-Kupper, M., Feroze-Merzoug, F. & Xi, Z. F. Secreted WNT antagonists as tumor suppressors: pro and con. *Front. Biosci.* **11**, 2093–2105 (2006).

20. Ranganathan, P. & Rangnekar, V. M. Regulation of cancer cell survival by Par-4. *Ann. NY Acad. Sci.* **1059**, 76–85 (2005).

21. Abdollahi, A. LOT1 (ZAC1/PLAGL1) and its family members: mechanisms and functions. *J. Cell. Physiol.* **210**, 16–25 (2007).

22. Nie, Y. *et al.* DNA hypermethylation is a mechanism for loss of expression of the HLA class I genes in human esophageal squamous cell carcinomas. *Carcinogenesis* **22**, 1615–1623 (2001).

23. Kenyon, K. *et al.* Lysyl oxidase and rrg messenger RNA. *Science* **253**, 802 (1991).

24. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl Acad. Sci. USA* **89**, 1827–1831 (1992).

Author Contributions S.G. and C.-M.V. prepared the retroviral library. C.G. and N.W. performed all experiments and data analyses. C.G., N.W. and M.R.G. designed the experiments, discussed the interpretation of the results and co-wrote the paper. All authors commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to M.R.G. (michael.green@umassmed.edu).

METHODS

Cell culture. NIH 3T3 (ATCC no. CRL-1658) and K:Molv NIH 3T3 (ATCC no. CRL-6361; referred to here as *K-ras* NIH 3T3) cells were maintained in DMEM supplemented with 10% FCS at 37 °C and 5% CO_2. For treatment with 5-aza, *K-ras* NIH 3T3 cells were treated with 10 μM 5-aza for 72 h.

shRNA screen. The mouse shRNA^{mir} library (release 2.16; Open Biosystems) was obtained through the University of Massachusetts Medical School shRNA library core facility. Ten retroviral pools, each comprising about 6,000 shRNA clones, were generated with titres of about 2.6×10^5 plaque-forming units ml^{-1}. These retroviral stocks were produced after co-transfection into the PhoenixGP packaging cell line (a gift from G. Nolan). *K-ras* NIH 3T3 cells (1.2×10^6) were transduced at a multiplicity of infection of 0.2 with the retroviral stocks in 100-mm plates, and two days later were selected for resistance to puromycin (1.5 μg ml^{-1}) for seven days. To isolate Fas-positive cells, 5×10^6 cells from each pool were incubated with an anti-Fas antibody (15A7; eBiosciences) followed by incubation with IgG-conjugated magnetic beads (Miltenyi Biotec), and Fas-positive cells were selected by using the Mini MACS magnetic separation system (Miltenyi Biotec) in accordance with the manufacturer's instructions. The selected Fas-positive cells were expanded and genomic DNA was isolated. To identify the candidate shRNAs, the shRNA region of the transduced virus was PCR-amplified (using primers PSM2-forward (5′-GCTCGCTTCGGCAGCA-CATATAC-3′) and PSM2-reverse (5′-GAGACGTGCTACTTCCATTTGTC-3′)) and cloned into pGEM-T Easy (Promega). An average of 30 clones were sequenced per pool (using primer PSM2-seq, 5′-GAGGGCCTATTTCCC-ATGAT-3′). Individual knockdown cell lines were generated by retroviral transduction of 0.6×10^5 *K-ras* NIH 3T3 cells with the respective shRNA. Individual shRNAs were either obtained from the Open Biosystems library or synthesized (see Supplementary Table 1).

Immunoblot analysis. To prepare cell extracts, *K-ras* NIH 3T3 knockdown cell lines were harvested seven days after retroviral transduction and selection with puromycin (1.5 μg ml^{-1}) and lysed by being boiled in 1×SDS sample buffer (Laemmli buffer) for 5 min. Proteins were resolved by 12% SDS–PAGE. Immunoblot analysis was performed with an anti-Fas (sc-716; Santa Cruz) or anti-p21 Ras (ab16795; Abcam) antibody to monitor the expression of *K-Ras* (as a loading control), and an appropriate horseradish-peroxidase-conjugated secondary antibody. Proteins were detected with SuperSignal West Pico Luminol/Enhancer Solution (Pierce).

Bisulphite sequencing. For each cell line six clones were sequenced. The regions analysed were amplified by nested PCR. The first round comprised 24 cycles at 94 °C for 1 min, 48 °C for 1 min 30 s, and 72 °C for 1 min. One-tenth of the product was used as substrate for the second round of PCR comprising 28 cycles at 94 °C for 1 min, 48 °C for 1 min 30 s, and 72 °C for 1 min. Primer sequences are provided in Supplementary Table 2.

Chromatin immunoprecipitation (ChIP) and methylated DNA immunoprecipitation (MeDIP). The following antibodies were used: anti-5-methylcytosine (ab1884; Abcam), anti-EZH2 (4905; Cell Signalling Technology), anti-CTCF (07-729; Upstate), anti-BMI1 (ab14389; Abcam), anti-DNMT1 (IMG-261A; Imgenex), anti-SIRT6 (ASB-ARP32409; Aviva Systems Biology), anti-TRIM37 (a gift from A. E. Lehesjoki), anti-TRIM66 (a gift from R. Losson), and anti-NPM2 (a gift from M. M. Matzuk). The anti-ZFP354B antibody was raised against a synthetic peptide corresponding to amino-acid residues 126–143 of the murine protein, and was affinity purified on a peptide coupled to agarose. For ChIP analysis of the *Fas* promoter, primer pairs located at the core promoter/TSS, about 1 kb upstream of the TSS or about 2 kb upstream of the TSS, were used for PCR analysis of the input and immunoprecipitated DNA samples. MeDIP and ChIP products were detected by autoradiography, or analysed by quantitative real-time PCR with Platinum SYBR Green qPCR SuperMix-UDG with Rox (Invitrogen). Calculation of fold differences was performed as described previously[25].

Quantitative real-time RT–PCR. Total RNA was isolated by using Trizol (Invitrogen) seven days after retroviral transduction and selection with puromycin. Reverse transcription was performed with SuperScript II reverse transcriptase (Invitrogen) in accordance with the manufacturer's instructions, followed by quantitative real-time PCR as described above. Primer sequences are provided in Supplementary Table 2.

Tumour formation assays. NIH 3T3, *K-ras* NIH 3T3 or *K-ras* NIH 3T3 knockdown cells (5×10^6) were suspended in 100 μl of serum-free DMEM and injected subcutaneously into the right flank of athymic Balb/c (*nu/nu*) mice (Taconic). Tumour dimensions were measured every three days from the time of appearance of the tumours, and tumour volume was calculated with the formula $\pi \times$ (length) $\times$ (width)^2/6. Animal experiments were performed in accordance with the Institutional Animal Care and Use Committee (IACUC) guidelines.

25. Pfaffl, M. W. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.* **29**, e45 (2001).

nature

LETTERS

Open-to-closed transition in apo maltose-binding protein observed by paramagnetic NMR

Chun Tang[1], Charles D. Schwieters[2] & G. Marius Clore[1]

Large-scale domain rearrangements in proteins have long been recognized to have a critical function in ligand binding and recognition, catalysis and regulation[1–5]. Crystal structures have provided a static picture of the apo (usually open) and holo usually closed) states. The general question arises as to whether the apo state exists as a single species in which the closed state is energetically inaccessible and interdomain rearrangement is induced by ligand or substrate binding, or whether the predominantly open form already coexists in rapid equilibrium with a minor closed species. The maltose-binding protein (MBP), a member of the bacterial periplasmic binding protein family[6], provides a model system for investigating this problem because it has been the subject of extensive studies by crystallography[7,8], NMR[9–11] and other biophysical techniques[11–13]. Here we show that although paramagnetic relaxation enhancement (PRE) data for the sugar-bound form are consistent with the crystal structure of holo MBP, the PRE data for the apo state are indicative of a rapidly exchanging mixture (ns to μs regime) of a predominantly (∼95%) open form (represented by the apo crystal structure) and a minor (∼5%) partially closed species. Using ensemble simulated annealing refinement against the PRE data we are able to determine a $\langle r^{-6} \rangle$ ensemble average structure of the minor apo species and show that it is distinct from the sugar-bound state.

The PRE is proportional to the $\langle r^{-6} \rangle$ average of the distance between a proton and the paramagnetic centre. Owing to the large magnetic moment of an unpaired electron, the PRE is detectable up to large distances. In a system comprising two or more species in rapid exchange, the observed transverse PRE rate, Γ_2, is a population-weighted average of the Γ_2 rates of the component species[14,15]. As a result, the PRE is exquisitely sensitive to the presence of minor species, even as low as ∼0.5–1%, provided that the distances between the paramagnetic centre and the protons of interest are much shorter in the minor species than the major one. In previous work we made use of this property of the PRE to detect transient intermediates involved in the target search process whereby a transcription factor locates its specific DNA target site[14] and to reveal semi-quantitatively the distribution of an ensemble of transient encounter complexes that facilitate stereospecific complex formation in protein–protein association[15]. Here we extend the use of the PRE to elucidate quantitatively the structure of a minor protein conformer in rapid exchange with the major conformer in the context of interdomain rearrangement. The ability to observe minor species in an equilibrium ensemble affords new insights into our understanding of weakly populated high-energy states of proteins.

MBP comprises amino-terminal and carboxy-terminal domains (NTDs and CTDs, respectively) connected by three linker segments (Fig. 1)[7,8]. Crystal structures of the apo and holo states of MBP reveal hinge-bending within the linker region that results in a ∼35° rigid

body domain reorientation on the binding of sugar (Fig. 1a)[7,8]. Solution dipolar couplings (RDC) for the apo and holo (malto-triose-bound) states agree well with the respective crystal structures[10], with RDC R-factors[16] of ∼14% (Supplementary Fig. 1a, b). Small-angle X-ray scattering (SAXS) data for apo MBP in solution are also consistent with the apo crystal structure (Supplementary Fig. 2a). However, the RDC and SAXS observables, unlike the PRE, are linear weighted averages of the species present in solution and are therefore insensitive to states with low populations (Supplementary Fig. 1c).

The paramagnetic probe used for PRE measurements was a nitroxide spin-label (2,2,6,6-tetramethyl-1-piperidinyloxy; TEMPO) conjugated through a maleimide group to engineered cysteine residues, D41C and S211C (one at a time), located in the NTD and CTD, respectively, at the periphery of the interface between the two domains. The D41C and S211C sites have previously been used for electron paramagnetic resonance studies[12]. The structure in the vicinity of the two mutations is the same in the apo and holo crystal structures[7,8]; NMR data indicate that these regions are rigid[10], and the spin-labels do not perturb either the backbone structure or ligand-binding characteristics (Supplementary Figs 3 and 4, respectively).

Transverse PRE rates, Γ_2, of the backbone amide protons (^1H_N) were measured for holo (matotriose-bound) and apo MBP (Fig. 2a, b for D41C; Supplementary Fig. 6 for S211C). Although the intradomain PREs for holo and apo MBP are similar (Supplementary Fig. 7), the interdomain PREs are clearly different (Fig. 2a, b, and Supplementary Fig. 6c, d). To ascertain the agreement between observed and calculated PREs quantitatively, ensemble paramagnetic probe simulated annealing[17] was employed with a six-conformer representation (Supplementary Fig. 8) to optimize the conformational space sampled by the probe, refining either against the intradomain
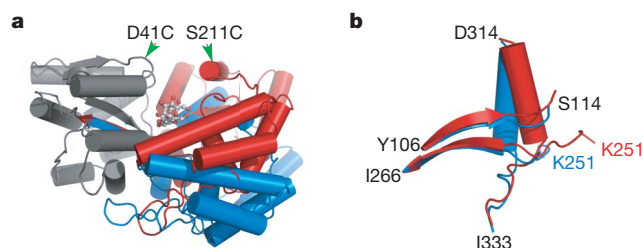

Figure 1 | Domain reorientation of MBP on binding maltotriose. a, View illustrating domain rearrangement in the crystal structures of apo (1OMP[7]) and holo (3MBP[8]) MBP. b, Close-up of the linker regions. The NTDs of apo and holo MBP are superimposed (grey); the conjugation sites (D41C and S211C) for the nitroxide spin-labels are indicated by the arrows in a, the CTDs and linker regions of apo and holo MBP are coloured in blue and red, respectively, and the maltotriose substrate is shown as balls and sticks.

[1]Laboratory of Chemical Physics, Building 5, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892-0520, USA. [2]Division of Computational Bioscience, Building 12A, Center for Information Technology, National Institutes of Health, Bethesda, Maryland 20892-5624, USA.

PREs or against the complete PRE data sets. For both holo and apo MBP, good agreement with the intradomain PRE data are obtained (PRE Q factors[15,17], Q_{PRE}, of 0.14 and 0.20, respectively, for D41C, and 0.25 and 0.22, respectively, for S211C; Supplementary Figs 5a, b and 6a, b).

For holo MBP, refinement against the intradomain PRE data predicts the interdomain Γ_2 rates well (Fig. 2a and Supplementary Fig. 5d). Paramagnetic probe refinement against all the PRE data yields Q_{PRE} values of 0.18 for D41C (Supplementary Fig. 5f, h) and 0.34 for S211C (Supplementary Fig. 6f, h) with correlation coefficients (r) of 0.98 and 0.91, respectively. Thus, the PRE data for holo MBP are consistent with a single species that corresponds to the holo MBP crystal structure (although in this instance the PRE data would not be sensitive to the presence of a small population of open form).

For apo MBP, the PRE profile for S211C is also compatible with the crystal structure of the apo state ($Q_{PRE} = 0.23$, $r = 0.96$; Supplementary Fig. 6a, c). However, the interdomain PRE profile for apo MBP D41C is inconsistent with the apo MBP crystal structure when

paramagnetic probe refinement is performed against either the intradomain PRE data alone (Fig. 2b and Supplementary Fig. 5a, c) or all the PRE data (Fig. 2c and Supplementary Fig. 5e, g). The discrepancies are characterized by much larger than expected interdomain experimental Γ_2 values for residues 150–157, 209–211, 226–232 and 339–348 in the CTD and residues 111–112 and 258–262 within the linker, which form a contiguous surface facing the NTD (Fig. 2b). The absence of any intermolecular PREs in a mixed sample comprising D41C nitroxide spin-labelled apo MBP at natural isotopic abundance and wild-type U-^{15}N/^{2}H-labelled apo MBP (Supplementary Fig. 9) excludes random intermolecular collisions (that is, a
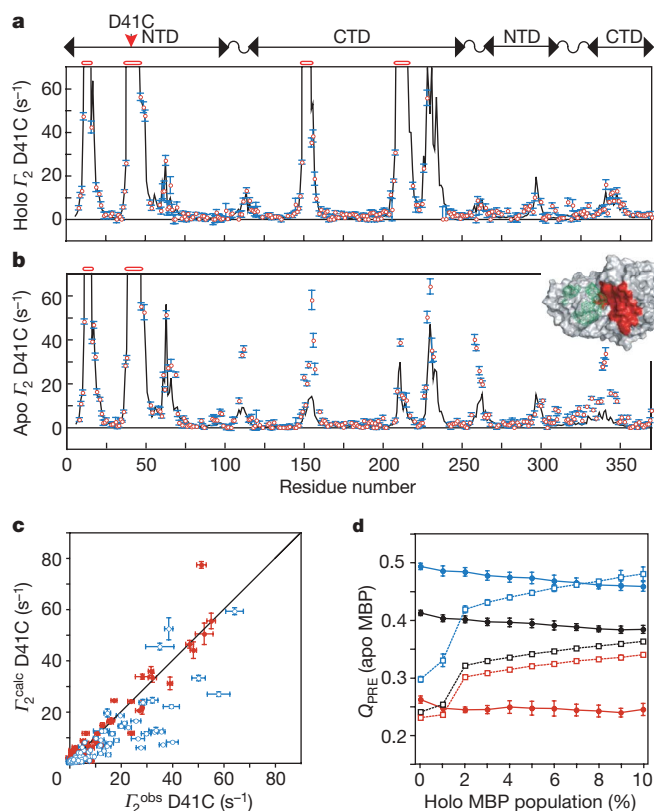

Figure 2 | PRE measurements on apo and holo MBP. a, b, Comparison of observed PRE profiles (red circles) with back-calculated values (averaged from 100 independent calculations, black lines) obtained by ensemble paramagnetic probe refinement[17] against the intramolecular PRE data for holo (**a**) and apo (**b**) MBP D41C. Residues whose ^{1}H$_N$/^{15}N cross-peaks are completely broadened out are denoted by open red bars. The inset in **b** shows a molecular surface of apo MBP, with the conformational space sampled by the nitroxide oxygen atom bearing the unpaired electron of the spin-label represented by a green mesh and the residues with PRE discrepancies, ($\Gamma_2^{obs} - \Gamma_2^{calc}$) > 10 s^{-1} for apo MBP coloured in red. In the schematic diagram above **a**, the wavy lines represent linker regions. **c,** Correlation between observed and calculated PREs for apo MBP D41C obtained with probe refinement against all the observed D41C PRE data. The intradomain and interdomain PREs are shown as red and blue circles, respectively. Overall $Q = 0.41$; overall correlation coefficient, $r = 0.88$; intradomain $Q = 0.26$; intradomain $r = 0.96$; interdomain $Q = 0.49$; interdomain $r = 0.85$. **d,** Q_{PRE} for the D41C (filled circles) and S211C (open squares) apo MBP PRE data for linear combinations of apo (1OMP[7]) and holo (3MBP[8]) MBP. Overall, intradomain and interdomain PREs are shown in black, red and blue, respectively. All error bars represent 1 s.d.
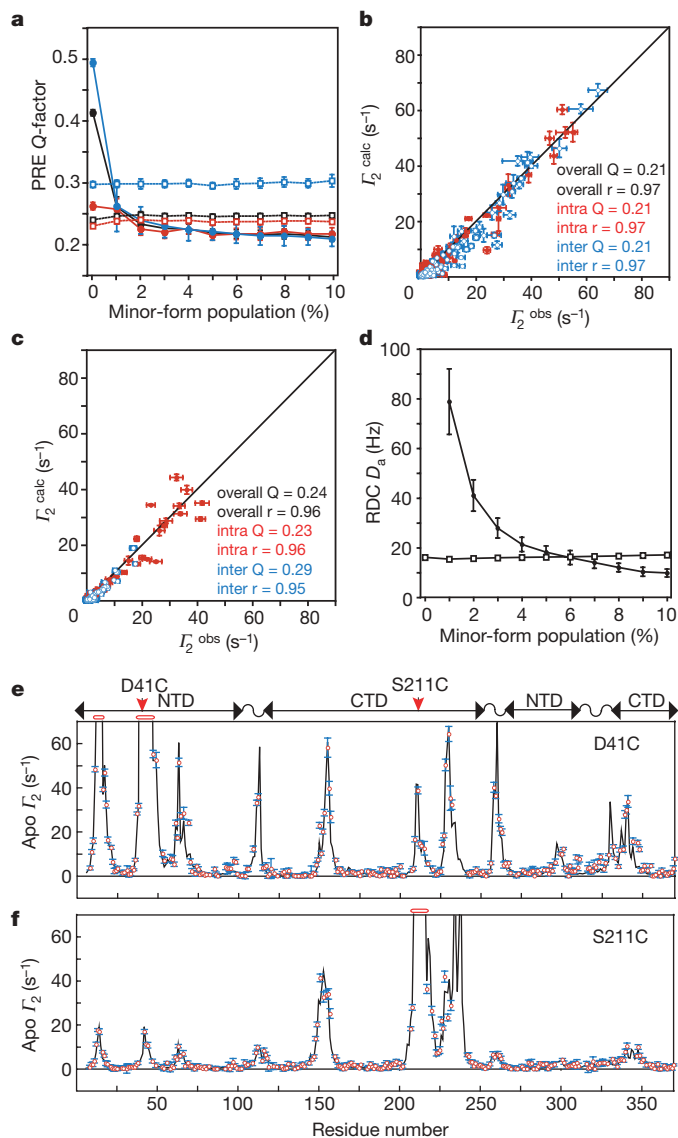
Figure 3 | Simulated annealing refinement of the minor species of apo MBP. a, Q_{PRE} as a function of minor-form population obtained by conjoined rigid-body/torsion-angle simulated annealing. Overall, intradomain and interdomain PREs are shown in black, red and blue, respectively; D41C, filled circles; S211C, open squares. Each data point is the average of 50 independent calculations. **b, c,** Correlation between observed and calculated PREs arising from D41C (**b**) and S211C (**c**) on introducing a minor species at an occupancy of 5%. **d,** Estimation of minor-species population from RDC data based on the variation of the calculated values of D_a^{NH} for the open (open squares) and minor (filled circles) forms of apo MBP obtained by SVD. The rhombicity is 0.29 ± 0.01 for the major form and 0.39 ± 0.12 for the minor form. **e, f,** Comparison of observed (red circles) and calculated (black lines) PRE profiles for D41C (**e**) and S211C (**f**) for a 5% minor species population. In the schematic diagram above **e**, the wavy lines represent linker regions. All error bars represent 1 s.d.

solvent PRE) or non-specific intermolecular interactions as a source of the discrepancies.

The simplest interpretation of the apo MBP PRE data—that the relative orientation of the NTD and CTD in solution is different from that in the apo MBP crystal structure—is readily excluded. Although a single conformation can be generated by rigid-body refinement[18] that satisfies the D41C and S211C PRE data simultaneously (Supplementary Fig. 10a, b), it is inconsistent with the RDC data (Supplementary Fig. 10c). Apo MBP in solution must therefore consist of a mixture of two (or more) states in rapid equilibrium with the predominant state corresponding to the apo MBP crystal structure. The relative orientation of the NTD and CTD in the minor species is distinct from that of the holo conformation because a linear combination of apo and holo states (with paramagnetic probe refinement) results in only a minimal decrease in Q_{PRE} for D41C and a substantial increase in Q_{PRE} for S211C (Fig. 2d). The latter excludes

an occupancy of greater than 1% for the closed holo conformation, in agreement with the extrapolated population of ~0.002% derived from thermodynamic data[11].

The structure of the minor species can be resolved by conjoined rigid-body/torsion-angle simulated annealing refinement[18] in which the PRE data for both D41C and S211C are fitted simultaneously to a two-member ensemble of major (open, 1OMP[7]) and minor species. The major form is held fixed in the apo crystal structure conformation, whereas the domains of the minor species are allowed to move as rigid bodies by giving the linker residues torsional degrees of freedom (see Methods). The overall and interdomain Q_{PRE} values for D41C decrease markedly on introduction of the minor species, even at a population as low as 1%, levelling off at a population of ~5%, whereas the good agreement with the PRE data for S211C remains unaffected (Fig. 3a). At a population of 5%, the overall Q_{PRE} is 0.21 for D41C (Fig. 3b) and 0.24 for S211C (Fig. 3c).
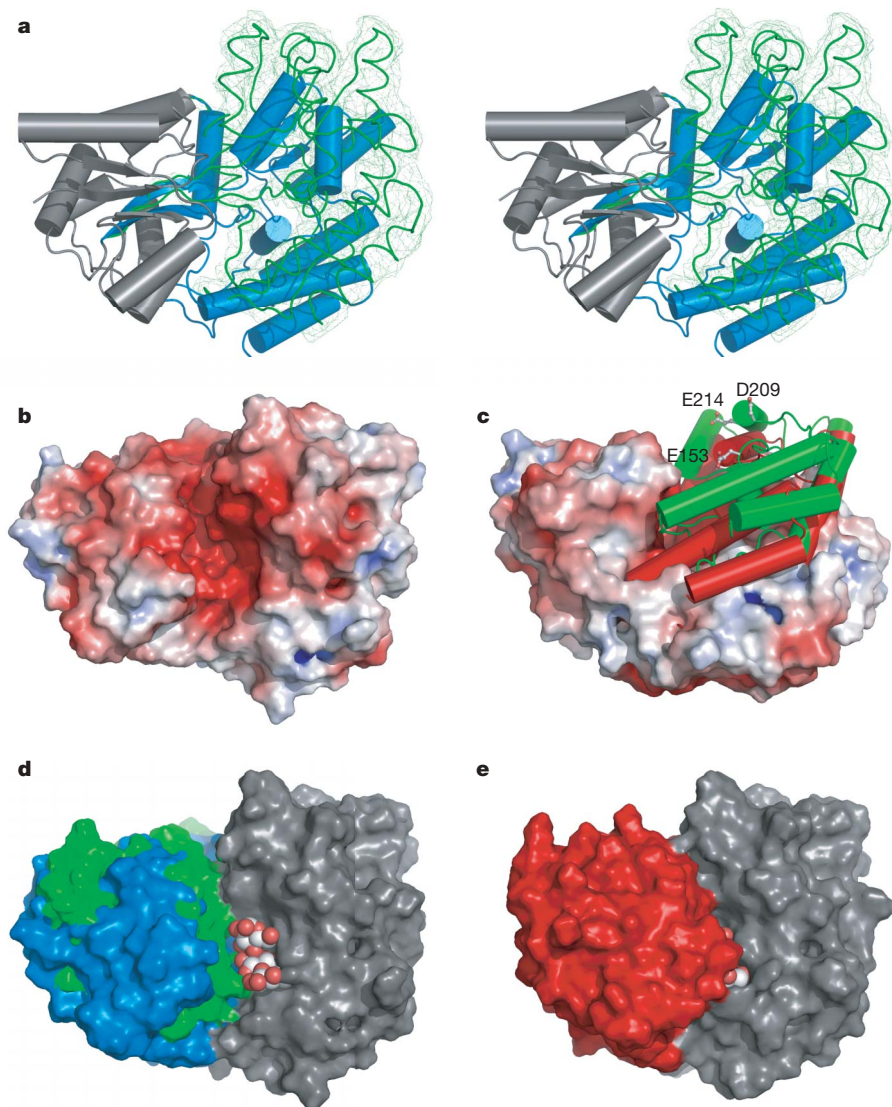

Figure 4 | Structure of the minor partly closed form of apo MBP.
a, Stereoview of the equilibrium mixture of the major open (blue cylinder) and minor closed (green smoothed backbone trace) forms of the CTD of apo MBP, with the NTDs of the two species superimposed and coloured in grey. The reweighted atomic probability map[25] for the backbone heavy atoms of the CTD in the minor form is displayed as a green mesh plotted at a threshold of 20%. **b**, Electrostatic potential surface[26] of the open form of apo MBP (1OMP[7]) drawn at $\pm 5kT$. **c**, Structural comparison of the CTD of the minor partly closed form of apo MBP (green cylinders) and holo MBP (red

cylinders; 3MBP[8]), with the open form of apo MBP shown as a molecular surface colour-coded as in **b** according to electrostatic potential. **d**, Molecular surface representation of the major open and minor partly closed species of apo MBP best-fitted to the CTD (grey) with the NTD (and linker) of the major and minor species shown in blue and green, respectively. A space-filling representation of maltotriose is modelled bound to the CTD[21]. **e**, Holo MBP in the same representation as in **d**, with the NTD shown in red; the substrate is buried in holo MBP.

Further, the Q_{PRE} values for the intradomain and interdomain D41C PRE data are comparable (Fig. 3b). The overall backbone coordinate precision for the minor species is 1.3 ± 0.4 Å, the backbone precision for the CTD when best-fitting to the NTD is 3.1 ± 1.0 Å, and the uncertainty in the relative orientation of the two domains is $10.5° \pm 4.0°$ (Fig. 4a and Supplementary Fig. 11). Representing the minor species by a three-conformer ensemble results in no significant change in Q_{PRE} relative to the single-conformer representation, indicating that a single minor species is sufficient to account fully for the experimental PRE data (Supplementary Fig. 12). Structures calculated against only the D41C PRE data cross-validate against the S211C PRE data and are essentially identical to those calculated from the combined D41C and S211C PRE data (Supplementary Fig. 13). The structures also cross-validate against PRE data measured on an EDTA-Mn^{2+} derivative of apo MBP D41C (Supplementary Fig. 14).

Because the magnitude (D_a^{NH}) of the RDC alignment tensor is determined largely by shape and charge distribution[19], and because D_a^{NH} is known to be very similar for the apo and holo states aligned by phage pf1 (Supplementary Fig. 1)[10] and can therefore also be assumed to be similar for the minor apo species calculated from the PRE data, the RDC data can be used to estimate the population of the minor species by using singular-value decomposition (SVD). Although the value of D_a^{NH} and the population of the minor species cannot be determined independently because they are anticorrelated (Supplementary Fig. 15), a plot of D_a^{NH} for the open and closed forms as a function of population (Fig. 3d) indicates that at a minor species population of 5–7% the major and minor forms have comparable values of D_a^{NH} (~16 Hz). An occupancy of ~5% for the minor species is also fully consistent with the SAXS data (Supplementary Fig. 2b).

The transition between major (open) and minor forms of apo MBP involves a hinge rotation of $33.3° \pm 6.7°$, in comparison with 35.2° between open apo and closed holo MBP. However, the apo minor and closed holo states of MBP are not the same and are related by a domain reorientation of ~18° accompanied by a ~6-Å translation (Fig. 4c). The timescale for the transition between major and minor species of apo MBP is fast, with an upper limit of ~20 μs derived from relaxation dispersion experiments[20] and a lower limit of ~20 ns, corresponding to the rotational correlation time, from {^{15}N}-^{1}H heteronuclear nuclear Overhauser enhancement data (see Methods and Supplementary Information).

The interface between the NTD and CTD is lined by negatively charged residues (Fig. 4b) that are responsible for an array of hydrogen bonds with the sugar substrate in holo MBP[8]. When the substrate is removed, the energy landscape is altered and access to the holo structure is energetically highly unfavourable as a result of potential electrostatic repulsion and lack of interdomain surface complementarity within the ligand-binding cleft that cannot be offset by bridging water molecules[11]. Unfavourable interdomain interactions in the minor form of apo MBP are circumvented by translation of the CTD out of the sugar-binding pocket, exposing the negatively charged residues Glu 153, Asp 209 and Glu 214 (Fig. 4c and Supplementary Fig. 16). The distance between the carboxylate of Glu 153 in the CTD to Glu 44, Glu 45 and Asp 14 in the NTD is decreased from ~15.3, ~15.8 and ~17.1 Å, respectively, in the open apo structure to ~7.2, ~9.1 and ~9.5 Å, respectively, in the closed holo structure, but to only ~13.5, ~14.8 and ~13.8 Å, respectively, in the minor apo structure. The latter is therefore best described as a partially closed apo conformation. The sugar-binding site in the closed holo conformation is inaccessible to incoming sugars because the NTD and CTD are closely apposed and engulf the bound sugar (Fig. 4e). In the open apo state, both the NTD and CTD sugar binding surfaces are accessible to bind sugar (Fig. 4d and Supplementary Fig. 17a, c). In the partially closed apo state, the NTD sugar-binding site is partly occluded but the CTD sugar-binding site is fully exposed (Fig. 4d, and Supplementary Fig. 17a, d and Supplementary Table 1). In this

regard we note the existence of a crystal structure of a MBP–malto-tetraitol complex in the fully open (apo) conformation with the ligand bound to the CTD[21].

The PRE observations reported here on apo MBP prove the existence of a long-postulated dynamic equilibrium between open and partially closed apo states, and in all likelihood reflect a general phenomenon of large-scale domain rearrangement observed in several multidomain proteins on ligand binding. The apo state exists as a rapid equilibrium of major and minor species, with the latter occupying a region of conformational space that is similar to but distinct from that occupied by the ligand-bound form. The presence of such minor species in the apo state may facilitate the transition to the holo structure that is rendered energetically accessible and stabilized by intermolecular interactions between the domains and the ligand. For MBP, the predominant fluctuations in the apo state do not involve the energetically disfavoured holo conformation and therefore the conformational change on sugar binding can be viewed as an example of induced fit.

METHODS SUMMARY

Sample preparation and NMR spectroscopy. Samples comprised 0.3 mM U-^{15}N/^{2}H-labelled MBP (D41C or S211C), which was either unmodifed (diamagnetic) or conjugated to maleimide-TEMPO (paramagnetic). ^{1}H$_N$-PRE data were acquired at 36 °C on a Bruker DRX-600 spectrometer with a cryogenic probe as described.[22] ^{1}D$_{NH}$ RDC data were obtained by taking the difference in ^{1}J$_{NH}$ couplings measured in aligned (~15 mg ml^{-1} phage pf1)[23] and isotropic (water) media.

Refinement and back-calculation of PREs. PREs were back-calculated by simulated annealing with Xplor-NIH[24] with the use of a six-conformer ensemble for the maleimide-TEMPO group to account for configurational space sampled owing to its flexibility[17]. Refinement of the structure of the minor form of apo MBP was performed by conjoined rigid-body/torsion-angle simulated annealing[18] subject to a target function comprising a PRE pseudopotential[17], a quartic van der Waals repulsion term to prevent atomic overlap, a multidimensional torsion-angle database potential of mean force, and appropriate torsion angle and hydrogen-bonding distance restraints within the linker. The calculated PRE for residue i of apo MBP is given by $\Gamma_{2,i}^{obs} = (1 - p)\Gamma_{2,i}^{open} + p\Gamma_{2,i}^{closed}$, where p is the population of the minor species (1–10%). The minor form of apo MBP was represented by a second conformer allowing rigid-body interdomain movement by giving the linker regions (residues 106–114, 252–265 and 314–333) full torsional degrees of freedom (with the exception of the helix, residues 315–326, whose backbone was also treated as a rigid body). The structure of the predominant open form was fixed during the calculation to that of the crystal structure of apo MBP (1OMP)[7]. The top 50 structures, ranked on the basis of PRE and repulsive van der Waals energies, were used for analysis. Agreement between observed and calculated PRE data are given by Q_{PRE} (Methods, equations (2) and (3))[15,17].

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

1. Anderson, C. M. & Zucker, F. H. Steitz, T. Space-filling models of kinase clefts and conformational change. *Science* **204**, 375–380 (1979).
2. Gerstein, M., Lesk, A. M. & Chothia, C. Structural mechanisms for domain movements in proteins. *Biochemistry* **33**, 6739–6749 (1994).
3. Gerstein, M. & Krebs, W. A. A database of macromolecular motions. *Nucleic Acids Res.* **26**, 4280–4290 (1998).
4. Hayward, S. & Berendsen, H. J. C. Systematic analysis of domain motions in proteins from conformational change: new results on citrate synthase and T4 lysozyme. *Proteins Struct. Funct. Genet.* **30**, 144–154 (1998).
5. Wolf-Watz, M. et al. Linkage between dynamics and catalysis in a thermophilic–mesophilic enzyme pair. *Nat. Struct. Mol. Biol.* **11**, 945–949 (2004).
6. Tam, R. & Saier, M. H. Jr. Structural, functional and evolutionary relationships among extracellular solute-binding receptors of bacteria. *Microbiol. Rev.* **57**, 320–346 (1993).
7. Sharff, A. J., Rodseth, L. E., Spurlino, J. C. & Quiocho, F. A. Crystallographic evidence of a large ligand-induced hinge-twist motion between the two domains of maltodextrin binding protein involved in active transport and chemotaxis. *Biochemistry* **31**, 10657–10663 (1992).
8. Quiocho, F. A., Spurlino, J. C., &. Rodseth, L. E. Extensive features of tight oligosaccharide binding revealed in high-resolution structures of the maltodextrin transport/chemosensory receptor. *Structure* **5**, 997–1015 (1997).

1081

9. Skynnikov, N. R. et al. Orienting domains in proteins using dipolar couplings measured by liquid-state NMR: differences in solution and crystal forms of maltodextrin binding protein loaded with β-cyclodextrin. *J. Mol. Biol.* **295**, 1265–1273 (2000).

10. Evenäs, J. et al. Ligand-induced structural changes to maltodextrin-binding protein as studied by solution NMR spectroscopy. *J. Mol. Biol.* **309**, 961–974 (2001).

11. Millet, O., Hudson, R. P. & Kay, L. E. The energetic cost of domain reorientation in maltose-binding protein as studied by NMR and fluorescence spectroscopy. *Proc. Natl Acad. Sci. USA* **100**, 12700–12705 (2003).

12. Hall, J. A., Thorgeirsson, T. E., Liu, J., Shin, Y. K. & Nikaido, H. Two modes of ligand binding in maltose-binding protein of *Escherichia coli*: electron paramagnetic resonance study of ligand-induced global conformational changes by site-directed spin labeling. *J. Biol. Chem.* **272**, 17615–17622 (1997).

13. Zhang, Y. et al. Model of maltose-binding protein/chemoreceptor complex supports intrasubunit signaling mechanism. *Proc. Natl Acad. Sci. USA* **96**, 939–944 (1999).

14. Iwahara, J. & Clore, G. M. Detecting transient intermediates in macromolecular binding by paramagnetic NMR. *Nature* **440**, 1227–1230 (2006).

15. Tang, C., Iwahara, J. & Clore, G. M. Visualization of transient encounter complexes in protein-protein association. *Nature* **444**, 383–386 (2006).

16. Clore, G. M. & Garrett, D. S. R-factor, free R and complete cross-validation for dipolar coupling refinement of NMR structures. *J. Am. Chem. Soc.* **121**, 9008–9012 (1999).

17. Iwahara, J., Schwieters, C. D. & Clore, G. M. Ensemble approach for NMR structure refinement against ^{1}H paramagnetic relaxation enhancement data arising from a flexible paramagnetic group attached to a macromolecules. *J. Am. Chem. Soc.* **126**, 5879–5896 (2004).

18. Clore, G. M. & Bewley, C. A. Using conjoined rigid body/torsion angle simulated annealing to determine the relative orientation of covalently linked protein domains from dipolar couplings. *J. Magn. Reson.* **154**, 329–335 (2002).

19. Zweckstetter, M., Hummer, G. & Bax, A. Prediction of charge-induced molecular alignment of biomolecules dissolved in dilute liquid-crystalline phases. *Biophys. J.* **86**, 3444–3460 (1994).

20. Palmer, A. G. NMR characterization of the dynamics of biomacromolecules. *Chem. Rev.* **104**, 3623–3640 (2004).

21. Duan, X. & Quiocho, F. A. Structural evidence for a dominant role of nonpolar interactions in the binding of a transport/chemosensory receptor to its highly polar ligands. *Biochemistry* **41**, 706–712 (2002).

22. Iwahara, J., Tang, C. & Clore, G. M. Practical aspects of ^{1}H transverse paramagnetic relaxation enhancement measurements on macromolecules. *J. Magn. Reson.* **184**, 185–195 (2007).

23. Clore, G. M., Starich, M. R. & Gronenborn, A. M. Measurement of residual dipolar couplings of macromolecules aligned in a nematic phase of a colloidal suspension of rod-shaped viruses. *J. Am. Chem. Soc.* **120**, 10571–10572 (1998).

24. Schwieters, C. D., Kuszewski, J. & Clore, G. M. Using Xplor-NIH for NMR molecular structure determination. *Progr. NMR Spectrosc.* **48**, 47–62 (2006).

25. Schwieters, C. D. & Clore, G. M. Reweighted atomic densities to represent ensembles of NMR structures. *J. Biomol. NMR* **23**, 221–225 (2002).

26. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl Acad. Sci. USA* **98**, 10037–10041 (2001).

METHODS

Sample preparation. The gene encoding MBP was subcloned from the pMAL-c2x vector (New England Biolabs) into pET11a, and U-^{15}N/^{2}H-labelled (or U-^{15}N/^{13}C/^{2}H-labelled) MBP was overexpressed in *E. coli* in minimal medium prepared in ^{2}H$_2$O with ^{15}NH$_4$Cl and ^{2}H$_7$-glucose (or ^{13}C$_6$/^{2}H$_7$-glucose) as the sole nitrogen and carbon sources, respectively, supplemented by ^{15}N/^{2}H (or ^{15}N/^{13}C/^{2}H) Isogro (Isotec). The D41C and S211C point mutations were introduced one at a time by using a QuikChange kit (Stratagene). Wild-type and Cys-mutant MBP were purified sequentially with DEAE, Superdex-75 and MonoQ columns (GE Healthcare). To remove any bound carbohydrate, the protein was denatured in 2 M guanidine hydrochloride (GuHCl) and passed through a G-50 desalting column pre-equilibrated with 2 M GuHCl[27]. The denaturant was immediately removed by desalting, thereby allowing the protein to refold. Nitroxide spin-labelling was performed by adding a roughly tenfold molar excess of 4-maleimide-TEMPO (4-maleimide-2,2,6,6-tetramethyl-piperidine-1-oxyl; Aldrich) to a solution of ~20 μM MBP Cys mutant (D41C or S211C) in 10 mM Tris-HCl pH 7.4 buffer. After incubation at 25 °C for 30 min, the reaction was quenched by the addition of 10 mM 2-mercaptoethanol. The conjugated protein was further purified on a MonoQ column, and the modification was confirmed by electrospray ionization mass spectroscopy. NMR samples comprised 0.3 mM U-^{15}N/^{2}H-labelled protein (0.3 and 0.7 mM for PRE and RDC measurements, respectively) in 10 mM Tris-HCl pH 7.4 buffer, in both the absence (apo) and presence (holo) of maltotriose (2 equivalents). All NMR measurements were performed at 37 °C with Bruker spectrometers equipped with *z*-gradient cryogenic probes. Backbone assignments were based on previously published work[10] and verified by TROSY-based[28] triple-resonance NMR spectroscopy[29] using a sample of U-[^{15}N/^{13}C/^{2}H]MBP in both the apo and holo (maltotriose-bound) states.

RDC measurements. ^{1}D$_{NH}$ RDCs were obtained by taking the difference in J$_{NH}$ couplings in aligned (~15 mg ml^{-1} pf1 phage[23]) and isotropic (water) media[30]. J$_{NH}$ couplings were measured from the F_1 splitting between the TROSY and semi-TROSY components of the ^{1}H$_N$/^{15}N cross-peaks acquired in an interleaved manner with a ^{1}H–^{15}N TROSY-based correlation experiment[31]. The RDC *R*-factor, which scales from 0% to 100%, is defined as

$$R_{dip} = \{\langle (D_{obs} - D_{calc})^2 \rangle / (2\langle D_{random}^2 \rangle)\}^{1/2} \quad (1)$$

where D_{obs} and D_{calc} are the observed and calculated values of the RDCs, respectively, and $(2\langle D_{random}^2 \rangle)^{1/2}$ is the expected root mean squared deviation if the vectors were randomly distributed. The latter is given either by $(2\langle D_{obs}^2 \rangle)^{1/2}$ or exactly by $[2D_a^2(4 + 3\eta^2)/5]^{1/2}$, where D_a is the magnitude of the principal component of the alignment tensor and η is the rhombicity[16]. The latter formulation is used when a single alignment tensor is required (that is, a single structure), and the former when two alignment tensors are needed (that is, for two structures). The alignment tensor(s) and calculated RDC values were obtained by SVD in Xplor-NIH[24] by using the structure coordinates and the observed RDC values. For SVD calculations incorporating the co-ordinates of the major and minor forms of apo MBP, with populations $(1 - p)$ and p, respectively, the calculated RDC value for residue i is given by $D_{calc}^i = (1-p)D_{major}^i + pD_{minor}^i$. The accession codes for the coordinates of apo and holo MBP are 1OMP[7] and 3MBP[8], respectively, and protons were added by using the HBUILD facility in Xplor-NIH[24].

PRE measurements. PRE ^{1}H$_N$-Γ_2 rates are given by the difference in R_2 relaxation rates between the paramagnetic and diamagnetic states of the protein[17]. Conjugation between the maleimide ring and the sulphur atom of a Cys residue results in two equally populated enantiomers, up and down configurations relative to the plane of the maleimide ring. R_2 rates were determined from a two-time point (2 and 6 ms) interleaved measurement using a two-dimensional ^{1}H–^{15}N transverse-relaxation optimized spectroscopy (TROSY)-based experiment, as described previously[22]. As R_2 relaxation rates were determined from the initial portion of the exponential decay, the observed R_2 rate is the average of the R_2 rates for the two enantiomers (which are expected to be very similar). Each maleimide-TEMPO adduct enantiomer was therefore represented by three conformers, and PRE ^{1}H$_N$-Γ_2 rates were back-calculated with a six-conformer (3×2) ensemble representing the Cys-maleimide-TEMPO group to account for the flexible linker consisting of three rotatable bonds[17]. (Note that a fourth rotatable bond between the TEMPO moiety and the maleimide group does not affect the position of the nitroxide oxygen atom bearing the unpaired electron.) Because the electron relaxation rate τ_s of the free radical is much longer than that of the protein rotational correlation time τ_r, the PRE correlation time τ_c ($= (\tau_r^{-1} + \tau_s^{-1})^{-1}$) was assumed to be the same as τ_r (18.6 ns), as reported previously[22,32]. The conformational space sampled by the six-conformer ensemble of the Cys-maleimide-TEMPO group conjugated to either D41C or S211C was optimized by simulated annealing in torsion-angle space with Xplor-NIH[24]

(100 structures calculated) to minimize the difference between observed and calculated ^{1}H$_N$-Γ_2 rates[17]. The target function also includes stereochemical terms and a quartic van der Waals repulsion term to prevent atomic overlap between maleimide-TEMPO and the protein; overlap between the members of the Cys-maleimide-TEMPO ensemble, however, is permitted because the six-member ensemble represents a distribution of states. Agreement between observed and calculated Γ_2 rates is given by the PRE *Q*-factor, Q_{PRE} (ref. 17):

$$\langle Q_{PRE} \rangle = \left[\sum_i \{ \Gamma_{2,i}^{obs} - \Gamma_{2,i}^{calc} \}^2 / \sum_i (\Gamma_{2,i}^{obs})^2 \right]^{1/2} \quad (2)$$

Ensemble refinement of major and minor forms of apo MBP. Refinement against the PRE data for apo MBP was performed by conjoined rigid-body/torsion-angle refinement[18,33] with the use of Xplor-NIH[24] subject to a target function comprising a PRE pseudopotential[17], a quartic van der Waals repulsion term to prevent atomic overlap[34], a multidimensional torsion-angle database potential of mean force[35], hydrogen-bonding distance restraints to preserve the integrity of the β-sheet portion of the linker, and backbone ϕ/ψ and side-chain χ torsion angle restraints for the linker residues to keep these torsion angles within the range of values adopted in apo and holo MBP. The coordinates of the NTD (residues 1–105 and 266–313) are held fixed, the CTD (residues 115–251 and 334–370) is treated as a rigid body, and the linker segments (residues 106–114, 252–265, 314–333) are given full torsional degrees of freedom with the exception of the helix (residues 315–326), which is treated as a rigid body with side-chain torsional degrees of freedom. For calculations comprising a mixture of major and minor apo MBP, a two-member ensemble refinement is used in which the major form is held fixed in the apo crystal structure form (1OMP[7]) while the CTD of the minor species (represented as a rigid body) is allowed to move relative to the NTD by allowing the linker residues to be flexible. Calculations were also performed in which the minor form was represented not by a single structure but by an ensemble of three structures. The conformational space of the paramagnetic probe, represented by the six-conformer ensemble for Cys-maleimide-TEMPO, is simultaneously refined, and its similarity between the major and minor species is enforced by a weak non-crystallographic symmetry potential. The calculated PRE for residue i is given by $\Gamma_{2,i}^{calc} = (1-p)\Gamma_{2,i}^{major} + p\Gamma_{2,i}^{minor}$ (ref. 15), with the minor-species population ranging from 1% to 10%. For each population, the top 50 converged structures, ranked according to the PRE and van der Waals repulsive energies, were used for further analysis. The calculated Γ_2 rates for residue i displayed in the figures in both the main text and Supplementary Information are the averages over the 50 final structures, $\langle \Gamma_{2,i}^{calc} \rangle$, and the PRE *Q*-factors reported are based on these values[15]:

$$\langle Q_{PRE} \rangle = \left[\sum_i \left\{ \Gamma_{2,i}^{obs} - \langle \Gamma_{2,i}^{calc} \rangle \right\}^2 / \left(\sum_i \Gamma_{2,i}^{obs} \right)^2 \right]^{1/2} \quad (3)$$

Relaxation dispersion experiments. ^{1}H $R_{1\rho}$ relaxation dispersion experiments were conducted at a ^{1}H spectrometer frequency of 600 MHz, as described[36]. No significant decrease in transverse ^{1}H$_N$ relaxation rates were observed over ^{1}H spin lock field strengths ranging from 2 to 12 kHz (12,600–75,400 radians s^{-1}). These data indicate that the exchange rate between the major and minor forms of apo MBP is faster than ~70,000 s^{-1}. No significant decrease in {^{15}N}-^{1}H heteronuclear nuclear Overhauser effects is observed within the linker region[10], giving a lower limit of ~20 ns, corresponding to the rotational correlation time of MBP.

SAXS measurements. SAXS data were collected at the BioCAT (beam line 18-ID) of the Advanced Photon Source with the incident photons at 12 keV and a sample-to-detector distance of 242.8 cm. The 110-μl samples were loaded into a 1.5-mm diameter quartz capillary thermostatically controlled at 22 °C. The data were collected in a series of 300-ms individual frames with a total exposure time of 9 s, with the sample flowing back and forth throughout the measurement to minimize possible radiation damage. Two-dimensional images accumulated on the charge-coupled device detector were corrected for the individual pixel-counting efficiencies and dark-current rates, after which radial integrations were performed to obtain the one-dimensional scattering profiles. The scattering data collected for the empty capillary, the capillary filled with the buffer and the capillary filled with the protein solution were normalized for the incident beam intensities and the respective transmissions. No inconsistencies were found by inspection of the individual successive frames, indicating the absence of radiation damage to the sample during the measurement and permitting averaging of all acquired frames. After subtraction of the empty capillary scattering from both the sample and the buffer data, scattering of the pure solvent was subtracted from the protein solution scattering by taking into account the protein volume (calculated with partial specific volume of 0.73 cm^3 g^{-1} for the protein). The data were collected at two concentrations, 5.0 and 2.5 mg ml^{-1}, and identical profiles of the data with the lowest *Q* in the two

data sets were observed, indicating the absence of any interparticle interference effect. The higher-concentration data were used for the subsequent analysis because of their higher signal-to-noise ratio. The data were fitted to the respective structures by using CRYSOL (with a solvent electron density of $0.3340 \, e \, Å^{-3}$)[37].

Structure analysis. Structures were displayed in PyMol[38], and electrostatic potentials were calculated with APBS[26].

27. Ganesh, C., Shah, A. N., Swaminathan, C. P., Surolia, A. & Varadarajan, R. Thermodynamic characterization of the reversible, two-state unfolding of maltose binding protein, a large two-domain protein. *Biochemistry* **36**, 5020–5028 (1997).
28. Peruvshin, K., Riek, R., Wider, G. & Wüthrich, K. Attenuated T_2 relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biologial macromolecules in solution. *Proc. Natl Acad. Sci. USA* **94**, 12366–12371 (1997).
29. Yang, D. & Kay, L. E. TROSY triple-resonance four-dimensional NMR spectroscopy of a 46 ns tumbling protein. *J. Am. Chem. Soc.* **121**, 2571–2575 (1999).
30. Bax, A., Kontaxis, G. & Tjandra, N. Dipolar couplings in macromolecular structure determination. *Methods Enzymol.* **339**, 127–174 (2001).
31. Kontaxis, G., Clore, G. M. & Bax, A. Evaluation of cross-relaxation effects and measurement of one-bond couplings in proteins with short transverse relaxation times. *J. Magn. Reson.* **143**, 184–196 (2000).
32. Hwang, P. M., Skrynnikov, N. R. & Kay, L. E. Domain orientation in β-cyclodextrin-loaded maltose binding protein: diffusion anisotropy measurements confirm the results of a dipolar coupling study. *J. Biomol. NMR* **20**, 83–88 (2001).
33. Schwieters, C. D. & Clore, G. M. Internal coordinates for molecular dynamics and minimization in structure determination and refinement. *J. Magn. Reson.* **152**, 288–302 (2001).
34. Nilges, M., Gronenborn, A. M., Brünger, A. T. & Clore, G. M. Determination of three-dimensional structures of proteins by simulated annealing with interproton distance restraints: application to crambin, potato carboxypeptidase inhibitor and barley serine proteinase inhibitor 2. *Protein Eng.* **2**, 27–38 (1988).
35. Clore, G. M. & Kuszewski, J. χ_1 rotamer populations and angles of mobile surface side chains are accurately predicted by a torsion angle database potential of mean force. *J. Am. Chem. Soc.* **124**, 2866–2867 (2002).
36. Suh, J.-Y., Iwahara, J. & Clore, G. M. Intramolecular domain–domain association/dissociation and phosphoryl transfer in the mannitol transporter of *Escherichia coli* are not coupled. *Proc. Natl Acad. Sci. USA* **104**, 3153–3158 (2007).
37. Svergun, D., Barberato, C. & Koch, M. H. J. CRYSOL: a program to evaluate X-ray scattering of biological macromolecules from atomic coordinates. *J. Appl. Cryst.* **28**, 768–773 (1995).
38. DeLano, W. L. *The PyMol Molecular Graphics System*, ⟨http://www.pymol.org⟩ (DeLano Scientific, San Carlos, CA, 2002).

nature jobs

Many governments strive to build and sustain a vibrant scientific culture, investing a significant part of their gross domestic product in research and development. Presumably their aim is to encourage scientific research that ultimately will benefit the state, the economy and society at large. They are also looking to create jobs. But there can be too much of a good thing.

Japan's bold initiative to produce 10,000 postdocs by 2000 was designed to create better science jobs and foster discovery. But it has run into a fairly big problem — there are now too many postdocs and not enough jobs (see page 1084). Back in 1995, the Japanese government recognized a need to increase the number of postdocs so that young investigators had more opportunities to conduct research on their own ideas, operating in a more independent fashion than had been the case previously. By 1999, a year earlier than projected, Japan had reached its target. By 2005, there were 15,000 postdocs, 61% of whom were in science and engineering. Permanent jobs in academia began to dry up.

US scientists, especially biomedical scientists, will be all too familiar with this chain of events. For years, US science associations and vocal members of the scientific community have recognized that there is a glut of postdocs and a shortage of entry-level academic jobs (see *Nature* **422,** 354–355; 2003).

Although Japan and the United States have different cultures, there are common themes that could resolve their postdoc problems. A 2005 US National Academies report called *Bridges to Independence* suggested some solutions. Postdocs should get as much funding as possible for their own research, and there should be limits to the amount of time they spend under their respective principal investigators. Once that time is elapsed, they should become a different category of permanent employee.

Nations, it seems, face common challenges when adjusting their science workforces to maximize innovation, discovery and sustainable research opportunities for talented scientists.

Gene Russo, acting editor of *Naturejobs*

SCIENTISTS TO SPARE

After boosting its number of postdocs by thousands, Japan is dealing with a major researcher glut. **Heidi Ledford** reports.

YUKIKO SHINYA

KIMIMASA MAYAMA/EPA/CORBIS

Toshiyuki Kobayashi was offered his first faculty position at the age of 24. He hadn't published a single paper, and was yet to earn his PhD. "I didn't feel pressure to find a job, because there were not many students with a PhD," says the University of Tokyo mathematics professor. "That meant that I could concentrate on a very difficult problem, and I had time to study areas not related to my research field. It was very good for my research in the long run."

That was 1987. Now Kobayashi has four PhD students and two postdocs of his own. The academic market has changed dramatically, he says. He has seen postdocs struggle to find their next job, and he worries that the pressure on young mathematicians to publish is so acute that some tackle only small problems. "Their papers are often not very deep," says Kobayashi.

In 1995, the Japanese government launched a programme intended to more than triple the ranks of Japanese postdocs by 2000. Japan reached its goal — 10,000 postdocs — by 1999, and the numbers continued to climb. By 2005, there were more than 15,000 postdocs, 61% of them in science and engineering. But academic job openings failed to keep up.

This has left the system scrambling to find career options for researchers. "The changes occurred in the mid-1990s and still our students can't leave the system," says Kobayashi. "I feel that when the government started this, it didn't think seriously about the future."

Postdocs are a relatively new addition to the Japanese scientific scene. In the traditional Japanese system, people with PhDs were typically hired as assistant professors and continued working in the lab where they did their thesis. These were stable positions, but carried little independence. "In Japan, the culture used to be one of conformity," says Kiyoshi Kurokawa, science adviser to Japan's former prime minister Shinzo Abe and a professor at the National Graduate Institute for Policy Studies in Tokyo. "You were not really being tested to be someone independent, you were still working as a subordinate of your preceptor." In 2004, the Japan Society for the Promotion of Science said that more than half of Japanese researchers had worked at only one institution during their scientific career.

Aiming for flexibility

The 10,000-postdoc programme was designed to make the position commonplace, allow young researchers to work independently and to aid flexibility so that they could switch labs. Another motive was that the number of PhDs awarded in Japan more than doubled from 1980 to 2002, leaving the system with more researchers than it could absorb. The postdoc programme just postponed the struggle for employment, say critics.

"Opportunities for getting a faculty position at universities or national institutions are not increasing," says Yoshinori Kumazawa, a materials scientist at Nagoya University. He says staff at other institutions have told him that budget cuts have left them unable to fill empty positions, much less take on new ones.

That doesn't mean the programme was a failure, says Ken-ichi Arai, professor emeritus of biochemistry at the University of Tokyo. But he is concerned that research funding continues to focus on 'big science' projects headed by established scientists; few initiatives

Managing the postdoc glut: Hitoshi Kikumoto (top) and Ken-ichi Arai.

create positions and funding for young researchers. "I'm generally positive about the past ten years of academic and industrial reforms in Japan," he says, citing recent increases in the numbers of women and foreign scientists. But a fundamental problem remains, he says: too many postdocs and too few positions.

Japan's postdocs are not alone in facing uncertain employment. A recent study by the Federation of American Societies for Experimental Biology noted that, although the number of US PhDs awarded in biology has increased from about 4,000 to more than 7,000 since 1981, the proportion of US biomedical science PhDs holding tenure or tenure-track positions 5–6 years after earning their PhDs fell from nearly 35% in 1981 to just over 20% in 2003. But there is a key difference: in the United States, the number of biomedical PhDs who take industrial jobs has also risen, and now encompasses 30% of the PhD pool. In Japan, only 0.2% of postdocs worked for private corporations in 2005. That amounted to 32 people.

Company reluctance

Japanese companies don't traditionally hire PhDs, preferring younger, less specialized researchers who are perceived as being more malleable, says Kurokawa. Raymond Price, now a manager at the biotech firm Acucela in Bothell, Washington, says that the Japanese drug firm he worked for in Osaka had only two other PhDs in a neuroscience research group of 20 scientists. And those two had earned their doctorates while working for the company, through the 'paper PhD' system, which allows industry scientists to submit their company research as their thesis — almost unheard of in the West. Of the entire staff of 120 researchers, fewer than five had joined the company with a PhD, Price says. Such hiring patterns have effectively closed off industry to those with an advanced degree (see pie chart). "Once you do a PhD and a postdoc, you don't really have a choice but to go to an academic institution," says Kurokawa.

Yet policy-makers are now looking to industry to provide jobs for postdocs. One option, says Hitoshi Kikumoto of the University of Tsukuba's Industrial Liaison and Cooperative Research Center, is to reshape graduate school curricula to prepare PhDs for work in industry. Japan's education ministry has established a programme to promote the diversification of science and technology career paths. It has provided funding to nine universities, two public research institutes and an

Raymond Price (below) notes that Japanese PhDs nearly always end up in academia —such as at the University of Tokyo (above).

academic society to come up with options for postdocs. Shinichi Kobayashi, who chairs such a committee at the University of Tsukuba, says that most programmes have focused on encouraging postdocs to find jobs in industry. But, he notes, the programmes are "very modest" and unlikely to find real solutions.

Overcoming company doubts will be an important first step, says Arai. "The structure of traditional Japanese industry is also vertical and hierarchical," he says, "and it has been difficult to employ someone with a PhD already in the middle stage of their career." The emerging biotech industry may provide opportunities, he says, but is still young and unpredictable. Kurokawa has been lobbying the companies themselves, arguing that their refusal to hire PhDs stifles innovation by limiting the influx of outside ideas. Yet it will take more time before the commercial sector adopts that sense of flexibility, he says. In the meantime, postdocs will have to wait. "It was a bit premature to rapidly increase this postdoc pool," says Kurokawa. "So they tend to suffer."
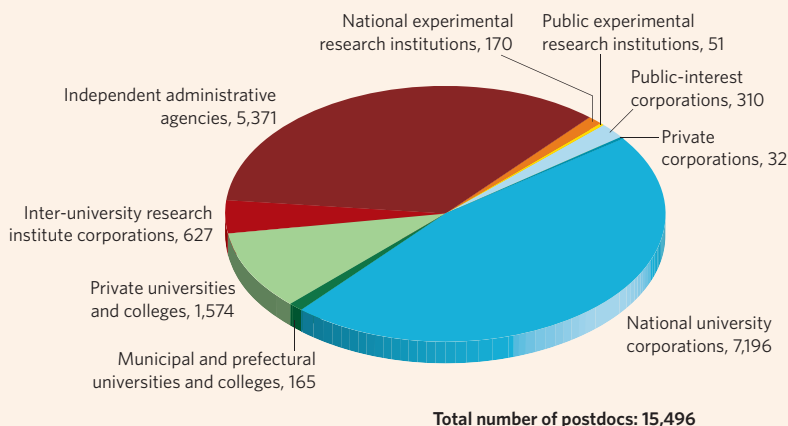
A survey of 83 Osaka University science and technology postdocs found that 85% of them were insecure about their careers and 60% no longer wanted to be postdocs (see *Nature* **447,** 1028; 2007). Short-term appointments leave postdocs little reprieve from job hunting, and the age limitations sometimes placed on postdoc positions add to the stress, notes Shoji Satoru, a nanotechnology postdoc at Osaka University. "I have to start job hunting again after a couple of years," he says. Meanwhile, he worries about what will happen when he reaches the age of 35 and loses eligibility for some postdoc positions.

Still, Satoru and others say they embarked on their scientific careers with their eyes open, having decided that the benefits of research outweigh the trade-offs of insecure employment. "All my colleagues are worried about it, but everyone still remains in this field," says Kyouhei Arita, a postdoc in a biochemistry lab at Kyoto University. Arita says that if he one day has postdocs of his own, he will advise them to think hard about their career decisions. "It is very hard to work as a postdoctoral scientist," he says.

Yet Kumazawa says postdoc concerns are reminiscent of those that scientists had when he was a graduate student: some felt that academic jobs were scarce then, too, he says. The problem was called 'the overdoctor issue'. "We occasionally had discussions about this issue, but no good remedy of our own," says Kumazawa. Scientists simply accepted it, he says. "They chose this career. They were happy and proud of being a scientist, even though they were unstable and poor."

Heidi Ledford writes for *Nature* from Boston.

NUMBER OF JAPANESE POSTDOCS BY ORGANIZATION, 2005

National experimental research institutions, 170

Public experimental research institutions, 51

Public-interest corporations, 310

Private corporations, 32

Independent administrative agencies, 5,371

Inter-university research institute corporations, 627

Private universities and colleges, 1,574

Municipal and prefectural universities and colleges, 165

National university corporations, 7,196

Total number of postdocs: 15,496

MOVERS

Nancy Andrews, vice-chancellor and dean, Duke University School of Medicine, Durham, North Carolina

2003–07: Professor of paediatrics, Harvard Medical School, Boston, Massachusetts
2003–07: Associate dean then dean, basic sciences and graduate studies, Harvard Medical School
1999–2003: director, Harvard-MIT MD-PhD programme, Harvard Medical School

As the new dean of a top US medical school, Nancy Andrews holds a position only a handful of women have achieved. While getting her first degree in molecular biophysics and biochemistry at Yale University, Andrews knew she wanted a research career, but she didn't contemplate medical school until she saw the opportunity to work at the interface of medicine and science. She completed her PhD in biology at the Massachusetts Institute of Technology (MIT) in Cambridge and her MD at Harvard. She chose paediatrics for her residency training in part because she enjoyed genetics, the basis of many childhood diseases.

Her path shifted towards administration once she became a faculty member at Harvard Medical School, as she hoped to help improve the Harvard-MIT MD-PhD programme. "It wasn't optimal for students," she says. The programme needed better integration between departments to speed up the time to graduation. Her former mentor at MIT, David Nathan, says that as soon as Andrews took over, the programme flourished — in part because she got to know each student individually. "She got in the trenches with the troops," he says.

Appointed dean for basic sciences and graduate studies at Harvard, Andrews relished problem-solving and helping to develop young careers. All the while, she has overseen a productive lab, working on the regulation of iron homeostasis and its relevance to disorders such as anaemia and haemochromatosis.

Over the past few years, she has been approached to take on a number of senior administration jobs, but none suited her until the offer came from Duke. "I was impressed by the culture and how easily innovation comes there," she says. What she didn't anticipate was the hoopla made of her appointment. "I was surprised, but ultimately glad the fuss was made to remind people that gender equality is still a problem," she says.

Nathan likens Andrews' personal style to that of Eugene Stead, an innovative former chief of medicine who made Duke's medical school a top programme in the 1960s, in part by tending to the needs of younger faculty members. Although Andrews will undoubtedly face the problems of every major medical school — recruitment and retention of top faculty members, attracting the best students and devising novel ways to teach — Nathan expects she will maintain the standard of excellence that was fostered by Stead.
Virginia Gewin

NETWORKS & SUPPORT

Hopes for growth in Spain

Despite the occasional high-profile success, Spain has not really reached its potential in terms of scientific output. Is the problem cultural or it is directly related to limited resources? Perhaps it's a bit of both.

Spain has more than 150,000 scientific researchers, according to statistics collected by Eurostat (http://epp.eurostat.ec.europa.eu) — but leading European research countries such as France and Britain have twice that number, and the United States has ten times as many. The gross domestic expenditure per researcher is half of that in France or Britain, and one-third of that in the United States. But this alone doesn't explain why Spain publishes so few papers in *Science* or *Nature*: 5 to 10 times fewer than France or Britain and 50 times fewer than the United States. It even publishes fewer than smaller countries such as the Netherlands or Switzerland.

One problem is that funding in Spain rarely provides for students or postdocs. The national funding agency, the CICYT, relies mostly on permanent staff to carry out the benchwork and each scientist is allowed to be full-time on only one project at a time. Competitive research requires substantially more resources. Regional funding may add a small yearly amount to the CICYT budget: European Union money could make a difference, but it reaches only a few exceptional projects and groups.

Spain should invest generous funding in a few outstanding institutes. It currently has some 'centres of excellence' — research facilities selected by the government. Unfortunately, these compete with other institutions for funding and have only limited additional resources to service and renew equipment and assist visiting scientists. They do not have privileged funding for students or technical personnel. Change does not look likely.

Even if Spain does invest in these centres, it must also address the problems documented in a recent study (A. Sibert *Labour productivity growth in the European Union* www.europarl.europa.eu/comparl/econ/emu/20061220/sibert_en.pdf): low productivity and weak collaboration among Spanish research groups. The country's scientists will have to earn the trust of tax-payers, funding agencies and potentially collaborative companies for Spanish science to reach its full potential. ■
José Prieto is staff researcher at the Institute of Optoelectronics Systems and Microtechnology, Polytechnic University of Madrid.

POSTDOC JOURNAL

Metamorphosis

Several of my colleagues are facing personal and professional transitions. They are starting their own labs, switching to jobs in industry or becoming first-time parents. As I watch them adapt to their new roles, I am reminded that I too will soon face change, as I'm two years into my postdoctoral fellowship. What next? Do I apply for a university lectureship or look for another postdoc? Do I leave the bench altogether and pursue an alternative career? What about my husband, who's also a postdoc, and his aspirations? And what about a family?

Having moved to a new country, a new field and even a new model organism, I feel I should be accustomed to change. After all, change is constant in science, where progress is made with imaginative ideas and innovative technology, not old assumptions and outdated techniques. Yet the prospect of another transition still elicits some anxiety.

Change challenges us to leave our comfort zones and to evolve. But transitions, while initially daunting, can be very rewarding. I have learned additional skills, used novel technologies, met interesting people, and gained new insights about my work and myself. I think it's especially true in the case of scientists that an ability not only to adapt to change but also to anticipate it will help the ambitious individual soar above even his or her own expectations. ■
Maria Thelma Ocampo-Hafalla is a research fellow at Cancer Research UK's London Research Institute.

The patter of tiny feet

A blast from the past.

Guy Riddihough

Gilliam Wilbert read the news article with something akin to shocked disbelief. "Cockroach Extinct: household vermin bites the dust. *Blattella germanica*, the last remaining extant species of cockroach, has finally gone the way of the rat…"

Gilliam dug out the pile of old Schmidt boxes his grandfather, long dead, had made as part of a collection for the now defunct local museum. Gilliam had smuggled them into his cramped studio apartment one evening when the surveillance cameras had gone on the blink. Every species of cockroach: it had been the old man's speciality — more than that, the love of his life.

"See here, Gilly. A death's head cockroach. What a beauty, eh?" The young Gilliam had stared in awe. "And this one: here, put your hand out." Gilliam loved his grandfather deeply, still, he had trembled a little as the old man plopped something concealed and wriggling into his hand. "A giant burrowing cockroach, biggest in the world." Gill had looked with fascination at the plump brown insect that filled his child's hand. It had squirmed and wriggled lazily, too fat to run away.

Gilliam hunted through the Schmidt boxes and found the selfsame species, labelled *Macropanesthia rhinoceros*. He wondered if this might be the very insect grandpapa had let him handle all those years ago. How the old man had rambled on, Gilliam sitting on his knee: "These beauties have been with us for a hundred million years, since the peak of the dinosaurs, Gilly. Tough as nails, live without food for a month, and then they can scrape a living off no more than the glue on the back of a postage stamp."

At university Gilliam had taken a course in entomology, but it was already a dead science by then: almost all living insect specimens existed in a dwindling number of zoos. He had been forced to change to biochemical nutrient engineering. With the human population then at several tens of billions and counting, there was no room for anything not linked to the streamlined and efficient production of human biomass. Most fauna and flora were functionally extinct with space-efficient genome banks taking the place of the last protected habitats.

Slowly, through the years, he came to look upon the little scurrying creatures

as tiny arthropod revolutionaries, a secret last stand against an increasingly sterile world. And now even the cockroach had succumbed to the ever-expanding plague of humankind.

Gilliam riffled through the little kit as he took the tram back to the trans-Pennine dormitory megalopolis. Several small plastic tubes; the miniature disposable autocycler block; the little packet of freeze-dried host eggs with the word 'arthro' scribbled in marker pen on it. A legacy from one of his entomology practical courses: Species Resurrection 101.

Outside, the aseptic plascrete urbanscape slid silently by, tidy and neat. And in the far corner of the carriage, unseen by Gilliam, a minute camera swivelled and focused.

Back in his cramped apartment Gilliam opened one of the Schmidt boxes and selected a sample at random; *Periplaneta americana*. His heart thudded in his chest — all biotech activity was strictly regulated — but his hesitation lasted no more than an instant. He crushed the insect's mummified body into one of the tubes and slowly worked his way through the handwritten instructions he had penned at university, years before.

Over the following weeks the tiny eggs swelled and became taut, glistening in the dim light of the insectarium hidden away in his closet. The pale white nymphs hatched and grew and darkened, prospering under Gilliam's loving hand. The nymphs moulted into pale white adults,

and they, too, darkened to a soft golden brown.

Gilliam had sometimes dreamed of waking up as a cockroach, lying on his back, his insectile legs waving in the air. He'd roll onto the floor and scuttle across to a dark crack in the wall, twisting and squirming his way through. Safe in the darkness he'd be gripped by a sense of utter, total freedom.

"You little creatures spanned the globe," he whispered to them, "east to west, north to south, everywhere that humans went. And now…" He was cut short by a pounding at the door. Startled, he knocked the insectarium over, accidentally spilling its contents.

The biocrimes interrogator clasped his hands together on the desk in front of him.

"Gilliam, I'm on your side, but this does not look good, any way you spin it. Cockroaches, for crying out loud. Cockroaches live on filth, Gilliam, and they carry all kinds of nasty diseases. They walk all over our food and we end up eating our own filth and catching those diseases. Did you know that cockroach droppings cause potentially fatal allergic reactions, Gilliam?"

They took him to a holding cell.

"It's worse than I thought, Gilliam." The biocrimes interrogator wrinkled his nose at the smell and appeared to inspect a fingernail with great interest. "Those 'roaches you released, they're resistant to goddam nearly everything. What the hell were you doing? The prosecution is pushing for the big one."

"The what?" Gillian asked, the words whistling through broken teeth.

"Bioterrorism. You synthesized a listed biohazardous species."

Gilliam looked down at his bruised and swollen hands, "I just wanted to bring them back to life."

They threw the book at Gilliam, and he spent many years in a cell only slightly smaller than his apartment. Indeed, he might have spent the rest of his days there had it not been for a plague that swept the planet clean of most of the rest of the human race. A plague — it was rumoured — borne on the feet of cockroaches. ∎

The author is presently locked in a deadly cosmic battle with *Nature* — as he is an editor at *Science*, but has declared a truce for the length of this story.

JACEY